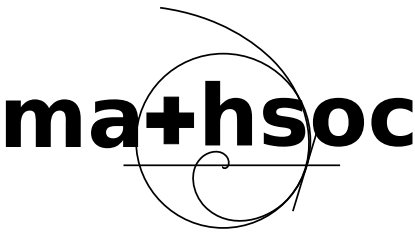
The Irish Undergraduate Mathematical Magazine

Issue 2: Summer 2013

The Irish Undergraduate
Mathematical Magazine
www.iumm.org
editors@iumm.org

Edited by James Fennell.
Assistant editor: Kieran Cooney.
Published August 2013.

The IUMM is a magazine run by and for undergraduate mathematicians in Ireland. We invite students and others to submit articles that are broadly mathematical and targeted at an undergraduate audience. See our website for more information on submitting.



The Summer 2013 issue was printed with the generous financial support of University College Cork Mathematics Society.

EDITOR'S NOTE

The introduction to the first issue of the *Irish Undergraduate Mathematical Magazine*, published 15 months ago, expressed the hope that the magazine could become a truly national publication. With the present issue containing articles from authors in six of the seven universities of the Republic of Ireland, it may be argued that that hope has been realized.

However, with all those who have been involved organizing the magazine in the United States next year, there will be no issue of the *IUMM* in 2014 unless there is a change of the guard. Any undergraduate student (or preferably students) in Ireland who feel up to the challenge of running the magazine are welcome to it. The only job requirement is a strong belief in the idea of communicating mathematics in an accessible way, and at least one student involved should have a reasonable knowledge of LaTeX typesetting. New editors will be given all of the existing typesetting files, and control of the domain name and email address.

The present magazine would, of course, not have been possible without the work of the contributors. Prof. Anthony O'Farrell generously volunteered his article to the magazine following his talk on the topic at the 2012 student conference in UCD; I am very grateful for his contribution. To the diverse group of undergraduate students who contributed to this issue: first my apologies, that the magazine took so long to go to print, but most of all my thanks for your hard work. I hope that you are happy with the finished product!

Finally, on a more personal note, I would like to thank to Dr. Tom Carroll and Niamh Tobin for each providing encouragement in their own way.

J.F.

AN INTRODUCTION TO THE CALCULUS OF VARIATIONS

KIERAN COONEY

University College Cork

In this article I hope to present an introduction to the calculus of variations that will not only whet the appetite of the reader for the subject, but that will also successfully describe an important mathematical tool that can help solve a very general range of problems. Most of us will be used to the "calculus of functions," such as taking derivatives or integrals of x^2 or $\sin(x)$, and from these gaining a much better understanding of the original functions. For many studying mathematics, the day you learn calculus is the first day of the rest of your life. In the calculus of variations, instead of studying functions of numbers, we study functions *of functions,* and find the analogous derivatives of these objects. This subject is far more general then its famous predecessor, but unfortunately not quite as developed.

1. FUNCTIONALS

Functions are very general things, so we need some rules to govern them before we start treating them like numbers. To begin with, we will consider functions from the real numbers to the real numbers. The set of continuous functions defined on a given closed interval $[a, b]$, which we call $\mathcal{F}$, is a vector space over the real numbers in the sense that addition and multiplication behave how we'd like them to. So we can add and subtract functions to get functions and we can multiply them by numbers to get functions. In general, we will consider $\mathcal{F}_n$, the set of all functions which map a given closed interval $[a, b]$ to $\mathbb{R}$ and which can be differentiated n times. Clearly $\mathcal{F}_n$ is also a vector space over the reals.

Now, we must introduce a notion of *distance* between our functions. We need this idea of distance, because to define differentiation we need to be able to take limits. Taking limits is equivalent to "getting close" to something, so naturally we need to know what "close" means. We call this notion of distance the *norm* of an element and we denote it by $\|x\|$. We want our notion of a norm to satisfy 3 conditions: for x and y vectors and α a scaler (real number) we want

1. $\|x\| = 0 \iff x = 0$.
2. $\|\alpha x\| = |\alpha| \, \|x\|$.
3. $\|x + y\| \leq \|x\| + \|y\|$.

Condition 3 is the familiar triangle inequality. It turns out that on $\mathcal{F}$ the most intuitive norm is also the most practical. For a given interval $[a, b]$ we define

$$\|f(x)\| = \max_{x \in [a,b]} |f(x)| \,. \tag{1.1}$$

We leave it as an exercise to the reader to show that this definition satisfies the above conditions.

This norm may appear overly simplistic. When I was studying the topic, I was more inclined to think that

$$\|f(x)\| = \int_a^b |f(x)|\,dx$$

would be a better norm as it seems to capture the "distance" of a function better. Indeed, it works perfectly also. However as we are really looking at the limiting behaviour of functions as opposed to their global behaviour, the above simpler definition suffices.

So now, using the first norm on the interval $[0, 2\pi]$, we have

$$\|A\sin(\omega x)\| = |A|.$$

This makes intuitive sense, as the distance of a function is simply the largest distance the function is away from the x-axis. We could extend this definition to $\mathcal{F}_n$, but there would be a problem. If we want a function to "get close" to another function, we want it to become very like that function in our limit. In our example above, the function $A\sin(\omega x)$ goes to zero as A goes to zero. But in $\mathcal{F}_1$ if we decrease A but increase ω, the norm of $A\sin(\omega x)$ will certainly go to zero but the derivative $A\omega\cos(\omega x)$ could stay constant or even diverge! In $\mathcal{F}_1$ we are studying functions *and* their derivatives so clearly this idea of closeness is wrong. To compensate for this, given a closed interval $[a, b]$ and an element $f(x)$ of $\mathcal{F}_n$ we define

$$\|f(x)\| = \sum_{i=0}^{n} \max_{x\in[a,b]} \left|f^{(i)}(x)\right|.$$

This definition considers the size of the function *and* its derivatives so that when two functions get close to each other they become alike in the sense we are interested in.

Now that we understand our "numbers" correctly, let's now examine our "functions". A *functional* is a function which takes a function and assigns to it a number. Let's take $\mathcal{F}_1$ for a given $[a, b]$. Then

$$J[f(x)] = f(a), \qquad J[f(x)] = \|f(x)\|,$$

$$J[f(x)] = \int_a^b f(x)dx, \text{ and} \qquad J[f(x)] = \frac{f'(a) + f'(b)}{2}$$

are all functionals. An interesting example of a functional is the total length of the graph of a function, which will be derived later in the article. We say a functional J on $\mathcal{F}_n$ is continuous at a point $\hat{y}$ (remember that $\hat{y}$ is a function now!) if for every positive ε there exists a positive δ such that

$$\|\hat{y} - y\| < \delta \Rightarrow |J[\hat{y}] - J[y]| < \varepsilon.$$

We can see why our norm definition was so important. If we had a functional that depended on the derivative of its argument but were still using the first norm (1.1) then the above definition of continuity would not make much sense.

We now introduce the *variation* of a functional. Set

$$\Delta J[h] = J[y + h] - J[y]$$

where y and h are both functions and J is our functional. Here h is behaving like our increment, or small difference, as it would in regular calculus. Suppose that we can write

$$\Delta J[h] = \varphi[h] + \varepsilon[h]$$

for some $\varepsilon[h]$ that goes to zero as $\|h\|$ goes to zero. Then we say φ is the *variation* of J. This is our "derivative". According to our definition, it behaves like a linear approximation, exactly as in the regular calculus case. In variational calculus, we do not "take derivatives" as much as we would in regular calculus, as the definition is too broad. Understanding what the variational is and that it exists though is key to the subject.

The most common type of functionals are of the form

$$J[y] = \int_a^b F(x, y(x), y'(x))dx = \int_a^b F(x, y, y')dx \tag{1.2}$$

where F is some real-valued function. These functionals are useful in that they can be constructed to depend on the local behaviour of the graph of a function at every point. For example,

$$J[y] = \int_a^b \sqrt{1 + y'(x)^2}dx$$

gives us the arc length of the graph of the function $y(x)$. You may ask the question why we say that y and y' are two independent parameters, as the derivative of a function is obviously defined by the function. This notation is convenient because it indicates explicitly what F depends on and also because we are often able to treat y and y' independently of each other in variation problems. In truth, the notation is just a matter of convenience.

2. THE EULER-LAGRANGE EQUATION

With all those definitions out of the way, let's start to apply what we have developed! Suppose that we are given a functional of the form (1.2) over $\mathcal{F}_2$ for some given interval $[a, b]$. Also, suppose we are restricted to functions with known end-points; that is, we know that $y(a) = A$ and $y(b) = B$ for some fixed A and B. We are looking for *extremals* of the functional J. An extremal is a function that is locally a maximum or a minimum of the functional J.

Clearly we require that the variation of J at y be zero if we want it to be an extremal. The proof of this statement is almost identical to the analogous regular calculus proof. We consider the effect of a small increment h on our functional:

$$\begin{aligned}\Delta J &= J[y+h] - J[y] \\ &= \int_a^b F(x, y+h, y'+h')dx - \int_a^b F(x, y, y')dx.\end{aligned}$$

Note that $h(a) = h(b) = 0$ to satisfy our boundary conditions. We now take a Taylor series expansion of F about y. There is no difficulty in doing this as F is just a real valued function. Then, taking terms up to first order gives

$$\Delta J = \int_a^b (F(x, y, y') + F_y(x, y, y')h + F_{y'}(x, y, y')h')dx - \int_a^b F(x, y, y')dx$$

where F_y and $F_{y'}$ denote the partial derivatives of F with respect to y and y' respectively. As we have neglected all but the linear terms of our Taylor expansion, ΔJ is in fact our variation. Using the linearity of integration and applying our extremal condition $\varphi[h] = 0$ we get

$$\varphi[h] = \int_a^b (F_y(x, y, y')h + F_{y'}(x, y, y')h')dx = 0.$$

Now we apply integration by parts to the second term in order to remove the dependence on h', remembering that $h(a) = h(b) = 0$. This gives

$$\begin{aligned}\int_a^b F_{y'}(x, y, y')h'dx &= F_{y'}(b)h(b) - F_{y'}(a)h(a) - \int_a^b h\frac{d}{dx}F_{y'}dx \\ &= -\int_a^b h\frac{d}{dx}F_{y'}dx\end{aligned}$$

and so then

$$\varphi[h] = \int_a^b h\left(F_y - \frac{d}{dx}F_{y'}\right)dx = 0.$$

We need this integral to vanish for any significantly small h. This implies that

$$F_y - \frac{d}{dx}F_{y'} = 0. \tag{1.3}$$

This equation is known as the *Euler-Lagrange equation.* This is *the* big equation in the calculus of variations.

As an example, let's try and find an extremal of the functional

$$J[y] = \int_0^{\pi/2} \left(y^2 - y'^2\right) dx$$

where $y(0) = 1$ and $y(\pi/2) = 0$. We have $F(x, y, y') = y^2 - y'^2$ so $F_y = 2y$ and $F_{y'} = -2y'$. Substituting this result into the Euler-Lagrange equation gives

$$2y - \frac{d}{dx}(-2y') = 0 \Longrightarrow y + y'' = 0$$

which is just our familiar ordinary differential equation for sinusoidal curves. So $y = A\cos(x) + B\sin(x)$. By substituting in our boundary conditions we see that $y = \cos(x)$ is our extremal. The questions remains as to whether $y = \cos(x)$ maximises or minimises the functional. It can be easily shown that it maximises it by looking at another function which satisfies the boundary conditions, say $y(x) = 1 - \frac{2x}{\pi}$, and showing that it has a smaller value.

After all those proofs and definitions, let me give you two important statements without proof which really illustrate the importance of the Euler-Lagrange equation. If we have a functional of the form

$$J[y_1, y_2, ..., y_n] = \int_a^b F[x,\ y_1,\ y_2, ..., y_n,\ y_1',\ y_2', ...,\ y_n']$$

where each y_i is a function of x with fixed end points, then at an extremal we must have

$$F_{y_i} - \frac{d}{dx} F_{y_i'} = 0 \ \text{ for all } i\,.$$

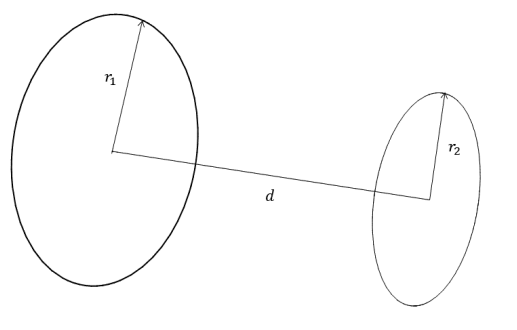
So finding a curve in $\mathbb{R}^n$ that maximises or minimises a functional reduces to just solving n different Euler-Lagrange equations! Secondly, the Euler-Lagrange equation is invariant under a change of variables. So if we are trying to find an extremal curve in 3-dimensions, we must solve 3 different Euler-Lagrange equations, but we can choose which co-ordinate system to use. We could use Cartesians, cylindrical polars, spherical polars, parabolic, hyperbolic, and so on. These two properties make this equation exceptionally general.

It should be clear to the reader that we haven't approached this topic from a strictly rigorous point of view, and there are some intricacies that we have simply ignored. This said, many satisfactory and rigorous resources on the topic may be found (see [1]).

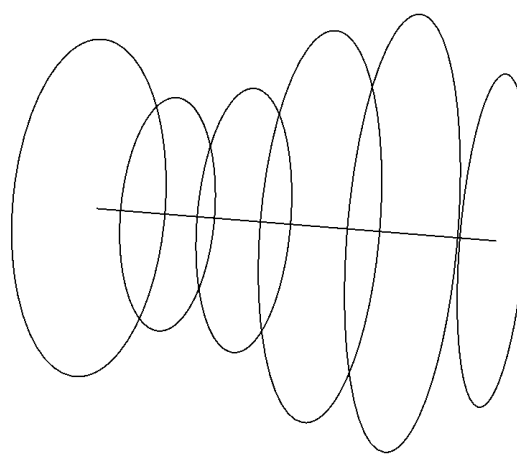
3. BUBBLES AND MINIMAL SURFACES

Mathematicians aren't happy unless there is a problem, so let's make one. I want to find the general shape of a bubble or soap film with a given boundary. For instance, what happens if I dip a helix into a soapy mixture and lift it out again. What shape will the soap film make? Intuitively, we know that soap

film is stretchy; i.e., there is a tension in the soap film that makes it contract as much as possible. This is the same thing as saying that a soap film tries to minimise its surface area as far as possible. In this way, we call bubbles *minimal surfaces*, in that they minimise surface area. The task remains to define the functional J which describes a given surface area.



Let's consider a more specific question. Suppose that I have two concentric, parallel rings which are a distance d apart and have radii r_1 and r_2 as shown in the diagram. If soap film was stretched between the two rings, what shape would it take? The shape is going to be a surface of revolution because of the axial symmetry of the construction. Suppose that we rotate a function $y(x)$ about the x axis. We can break up the graph into a sequence of co-axial rings. The perimeter of each ring will then be $2\pi y(x)$. Notice that if we take the limit of this sequence we must also account for the local behaviour of the graph $y(x)$. What we actually need is the local length of the graph $y(x)$. We presented an equation for this earlier, now let's justify it.



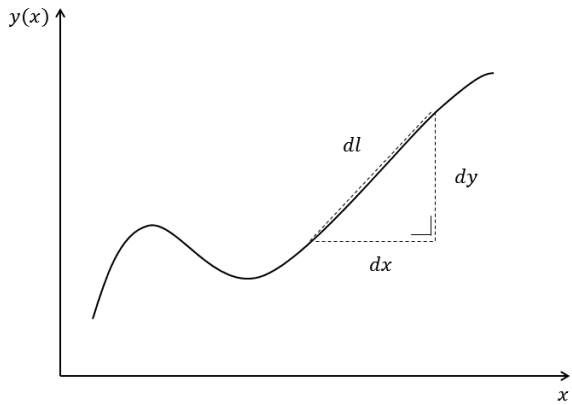
In the below diagram we have an arbitrary function $y(x)$. Here dl is the length of an infinitesimal part of the graph. By Pythagoras' Theorem we have

$dl^2 = dx^2 + dy^2$. Factoring out a dx^2 and taking square roots we see that

$$dl = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

giving us the above equation for the length of the curve y, denoted by $L[y]$:

$$L[y] = \int_a^b \sqrt{1 + y'^2}\, dx.$$



From this functional, if you wished, you could use the Euler-Lagrange equation to show that the straight line is indeed the shortest path between two points on the plane! Going back to the surface area problem, the local area of a surface of revolution of y is the perimeter of the loop times the local length of the graph. If we denote the surface area by $S[y]$ then

$$S[y] = \int_a^b 2\pi\, y\, \sqrt{1 + y'^2}\, dx.$$

Now that we have our functional, we may apply the Euler-Lagrange equation. As the S is independent of x, the Euler-Lagrange equation can be simplified ([1]) to

$$F - y' F_{y'} = C.$$

So in this instance the Euler-Lagrange equation becomes

$$y\sqrt{1 + y'^2} - y\frac{y'^2}{\sqrt{1 + y'^2}} = C.$$

We multiply across by $\sqrt{1 + y'^2}$ and factorise,

$$y(1 + y'^2 - y'^2) = y = C\sqrt{1 + y'^2},$$

re-arrange terms,

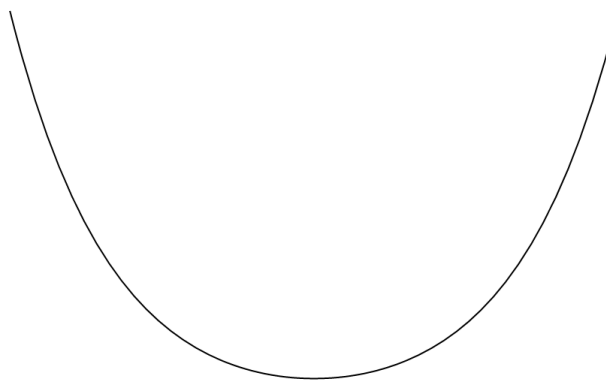
$$y' = \sqrt{\frac{y^2 - C^2}{C^2}},$$

and separate variables to obtain

$$dx = \frac{C\, dy}{\sqrt{y^2 - C^2}}.$$

An integration yields

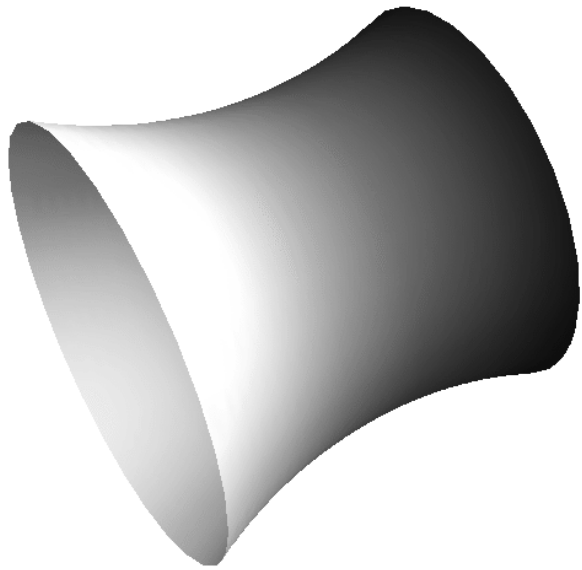
$$x + C_1 = C\, \ln\left(\frac{y + \sqrt{y^2 - C^2}}{C}\right).$$



The solution for this turns out to be a coshine function, defined by

$$\cosh(x) = \frac{1}{2}(e^x + e^{-x}).$$

This curve is know as a *catenary*. It is a very familiar shape because any time you hang a chord or a chain between two points, the resulting shape is a catenary (guess what, we can prove this using variational techniques). It is also used in constructing bridges and structural arcs as it supports weight very efficiently. We define the coshine function by If you rotate a catenary about its x axis then the resulting shape is a *catenoid*. It is the surface a soap film takes between two circular rings, and is the solution to our problem. To determine the values of our constants of integration C and C_1 we simply insert our boundary conditions.



The problem of finding a minimal surface given its boundary is known as *Plateau's problem*. Using a knowledge of differential geometry, we can find an equation that describes these surfaces (the surfaces turn out to be those with a mean curvature of zero). This is a very current field as the use of computer graphics has helped stimulate it enormously. In the 1970's, a Brazilian graduate student Celso José da Costa apparently solved the equations for a unique minimal surface in his sleep! The surface is known as *Costa's surface* and is very interesting in that it has a hole in it. One may find many inspiring images of these surfaces online, for instance at [4].

All this said about bubbles, we never explained why the regular ones are spheres! These bubbles are boundary less, so we need "another" boundary condition. We know that if there is a volume of air trapped inside the bubble, it must stay there. So the total volume enclosed by the bubble is fixed. As we know the bubble tries to minimise it's surface area, we want to find the shape that has the least surface area for a given volume. Intuitively, this is the sphere. And guess how we would show this is true? (I'll give you a hint: I'm writing an article on it.)

We may also extend (or contract) the idea of minimal surfaces to minimal curves or *geodesics.* These are curves which when restricted to a given surface minimise the distance locally. If we attempt to do geometry on surfaces, these are analogous to our lines in plane geometry. Again, we may use differential geometry and variational techniques to derive the equations that govern them.

4. THE BRACHISTRONE CURVE

We now deal with another variational problem, one of the first actually. Imagine that we are given two points in space and that we want to connect the two

points A and B by a curve. If we drop a ball on the curve at A, the ball will roll to B. We want to find the curve such that the ball will get from A to B in the least amount of time. This curve is known as the *brachistrone curve.* You may be tempted to say the curve will be the straight line, but it is not. We shall see why in a moment. The general velocity of a particle, by definition, is given by

$$v = \frac{ds}{dt} = \sqrt{1 + y'^2}\frac{dx}{dt},$$

where the second equality follows as before. We also know from conservation of energy ([3]) that $E = mgh + mv^2/2$ is constant. We will set $h = 0$ at the point A. Remembering that the ball will start from rest, set $v = \sqrt{2gy}$. So

$$\sqrt{2gy} = v = \sqrt{1 + y'^2}\frac{dx}{dt} \Rightarrow dt = \frac{\sqrt{1 + y'^2}}{\sqrt{2gy}}dx$$

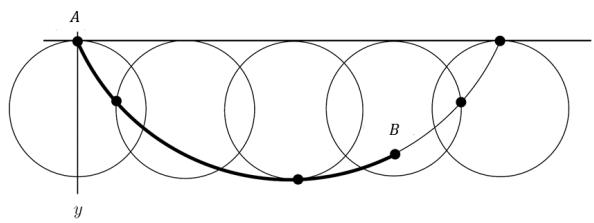
and the total time taken is given by the functional

$$T[y] = \int_a^b \frac{\sqrt{1 + y'^2}}{\sqrt{2gy}}dx$$

By applying the Euler-Lagrange equation again and solving for it as we did for the catenary, we obtain a family of solutions of the form

$$x(\theta) = \frac{1}{2C_1^2}(\theta - \sin(\theta)) + C_2, \qquad y(\theta) = \frac{1}{2C_1^2}(\theta - \cos(\theta)).$$

These curves are known as *cycloids*. They are the result of rolling a circle on a straight line and tracing out the path of a point on the circumference. They work because there is a steep initial slant to build up velocity and a gentle evening out to conserve it. There is another problem very similar to this one, the *tautochrone* problem. In this instance we wish to find a curve such that a ball will reach the bottom of it at some after some constant time T regardless of where it is dropped on the curve. Interestingly, this curve turns out to be the cycloid again!



5. ODDS AND ENDS AND THE END

The Calculus of Variations can be used in a wide range of topics, but excels particularly in physics (see [3]). So in the following paragraph, let me blurt out what else the calculus of variations allows us to do.

If we let V be the potential energy of a system and T be the kinetic energy of a body at a point in the system, then we define the *Lagrangian* by $L = T - V$. It turns out (see [2]) that Newton's laws of physics are equivalent to saying that the Lagrangian is minimised. This formulation is known as *the principle of least action. Fermat's principle* states that light will travel in such a way as to minimise the time of it's journey. Einstein's theory of relativity states that we *do not* live in a Euclidean universe and that if nothing gets in our way we travel along the geodesics of the universe. We may use variational techniques in electrodynamics to derive the equations of motion. Finally, Feynman took integrals over functionals instead of derivatives, and used this work to re-phrase quantum mechanics.

I hope I have opened your mind to a genuinely interesting field, and that I send you forth unto the world knowing how to use (or at least appreciate)

$$F_y - \frac{d}{dx} F_{y'} = 0.$$

BIBLIOGRAPHY

[1] I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Dover Publications, 2000.

[2] Tom W B Kibble and Frank H Berkshire. *Classical Mechanics*. World Scientific Publishing Company, 2004.

[3] Michael Mansfield and Colm O'Sullivan. *Understanding Physics*. Wiley, 2011.

[4] Paul Nylander. http://www.bugman123.com. Accessed: 2013-03-25.

INTEGRAL BINOMIAL COEFFICIENTS

PROF. ANTHONY O'FARRELL

National University of Ireland, Maynooth

1. INTRODUCTION

The binomial coefficients are defined by

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!},$$

for nonnegative integer k and any α. Usually, α is a real or complex number, but the definition makes sense if α belongs to any field of characteristic zero. The following is well-known:

Theorem 1.1. *The binomial coefficients $\binom{n}{k}$ are positive integers, for integers n, k with $0 \le k \le n$.* □

The usual proof uses the Law of Pascal's Triangle, and induction.

The binomial coefficients $\binom{r}{k}$, with rational r, occur in the Maclaurin series expansion of $(1+x)^r$ (convergent for real or complex x with $|x| < 1$). For instance,

$$\sqrt{1+x} = \sum_{k=0}^{\infty} \binom{1/2}{k} x^k.$$

Calculating a few terms, one finds that the series begins

$$1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{64}x^4 \cdots .$$

The coefficients are not integral (or nonnegative), but when common factors are cancelled (i.e. they are expressed in *reduced form* m/n, with $m \in \mathbb{Z}$, $n \in \mathbb{N}$, and $\gcd(m,n) = 1$), it is remarkable that only powers of 2 occur in the denominators. This is not an accident: the pattern continues forever. We have the following, slightly less well-known result:

Theorem 1.2. *Let $r \in \mathbb{Q}$ and $0 \le k \in \mathbb{Z}$. Suppose that $r = m/n$ in reduced form. Then the binomial coefficient $\binom{r}{k}$ has reduced form s/t, where t is a product of powers of primes that divide n.*

For instance, in the expansion of $(1+x)^{\frac{5}{6}}$, the coefficients all take the form $s/(2^a 3^b)$, for some $s \in \mathbb{Z}$.

The theorem may be proved using elementary number theory, for instance by reducing it to the statement that if $d, k \in \mathbb{N}$ and r is the largest factor of $k!$ prime to d, then r divides the product of the terms of each k-term arithmetic progression of integers having step d.

The purpose of this note is to give a very short soft proof of Theorem 1.2, by using a little analysis. Specifically, the proof uses the field $\mathbb{Q}_p$ of p-adic numbers. For the benefit of readers who have not met these numbers, we give a short introduction in the next section, and then give the proof in the final section.

2. THE P-ADIC NUMBERS

For a prime p, the *p-adic valuation* of a rational number is defined by setting $\|0\|_p = 0$ and

$$\left\| \frac{\pm r p^n}{s} \right\|_p = p^{-n}$$

whenever $r \in \mathbb{N}$, $s \in \mathbb{N}$, and $n \in \mathbb{Z}$ with $\gcd(r, p) = \gcd(s, p) = 1$. For instance,

$$\|300\|_2 = \frac{1}{4}, \ \|301\|_2 = 1, \text{ and } \left\| \frac{1}{300} \right\|_2 = 4.$$

Thus some numbers that have large absolute value have small valuations, and vice-versa. Also, numbers that have small valuations with respect to one prime may have large valuations with respect to another.

The *p-adic metric* on the set $\mathbb{Q}$ is defined by setting the distance between two rationals a and b equal to $\|a - b\|_p$. You can verify easily that this does, indeed, define a metric. In particular, the triangle inequality follows from a stronger form known as the *ultrametric* inequality:

$$\|a - b\|_p \leq \max\{\|a - c\|_p, \|c - b\|_p\}.$$

The space $\mathbb{Q}_p$ of *p-adic numbers* is the completion of $\mathbb{Q}$ with respect to the p-adic metric. It is a complete metric field, i.e. the field operations are continuous. One can show (although we do not need this for the proof below) that $\mathbb{Q}_p$ has the same cardinality as $\mathbb{R}$, and that it is locally-compact and totally-disconnected.

The closure of $\mathbb{Z}$ in $\mathbb{Q}_p$ is denoted $\mathbb{Z}_p$, and called the set of *p-adic integers*. It is a compact, totally-disconnected metric space, and an integral domain, and $\mathbb{Q}_p$ is its quotient field.

From the point of view of number theorists, there is little to choose between $\mathbb{R}$ and any of the $\mathbb{Q}_p$. They are all more-or-less equally-interesting ways to complete the set of rationals. For instance, if one is interested in solving a Diophantine equation such as $x^3 + y^3 = z^3$ for integers, then it is necessary that the equation have a solution in each $\mathbb{Z}_p$ and in $\mathbb{R}$. For some equations, the converse holds — such a result is called a "Hasse Principle".

Each infinite series of the form

$$\sum_{n=0}^{\infty} a_n p^n$$

with $a_n \in \mathbb{Z}$ is convergent in p-adic metric, and so represents some p-adic integer. For instance, in 2-adic metric we have the formula

$$1 + 2 + 4 + \cdots + 2^n + \cdots = -1,$$

which may be found in Euler's work. More generally, for any prime p,

$$(p-1) + (p-1)p + (p-1)^2 p + \cdots = -1$$

in p-adic metric. From this we deduce that every p-adic integer is the limit of a sequence of *positive* integers.

A non-integral rational number may be a p-adic integer. For instance,

$$1 + 3 + 3^2 + 3^3 + \cdots = -\frac{1}{2}$$

in 3-adic metric. More generally, it is not hard to see that a rational number r with reduced form m/n belongs to $\mathbb{Z}_p$ if and only if p does not divide n.

3. THE PROOF

Theorem 3.1. *If $p \in \mathbb{N}$ is prime, $a \in \mathbb{Z}_p$ and $0 \geq k \in \mathbb{Z}$, then $\binom{a}{k} \in \mathbb{Z}_p$.*

Proof. Fix $k \in \mathbb{Z}$, $k \geq 0$. The function

$$f : x \mapsto \binom{x}{k}$$

is a polynomial with coefficients in $\mathbb{Q}$, and hence it is continuous, as a function from $\mathbb{Q}_p$ into $\mathbb{Q}_p$. (This just depends on the fact that $\mathbb{Q}_p$ is a metric field.) Choose a sequence $(a_n)_{n=1}^{\infty} \subset \mathbb{N}$ with $a_n \to a$ in p-adic metric. Then $f(a_n) \in \mathbb{N} \subset \mathbb{Z}_p$, and hence $f(a) = \lim_{n\to\infty} f(a_n) \in \mathbb{Z}_p$, since $\mathbb{Z}_p$ is closed. □

We remark that a rational number r is an integer if and only if $r \in \mathbb{Z}_p$ for each prime p, and so this theorem may be regarded as a "local version" of Theorem 1.1. The proof shows that the local version follows at once from Theorem 1.1, and a simple bit of topology.

Proof of Theorem 1.2. Let $r = m/n$, k, and $\binom{r}{k} = s/t$ be as in the statement. Suppose a prime p divides t. If p does not divide n, then $r \in \mathbb{Z}_p$, so $s/t \in \mathbb{Z}_p$, which is false. Thus each prime that divides t divides n. □

4. POSTSCRIPT: AN "ELEMENTARY" PROOF

A special case of Theorem 1.2 is that $\binom{n}{k}$ is integral for all *negative integers* n and $0 < k \in \mathbb{N}$. It is possible to prove this directly by extending Pascal's triangle backwards and using induction (we show a tilted version, in which the row

corresponding to $n \in \mathbb{Z}$ has the entries $\binom{n}{k}$, $(k \in \mathbb{N})$ beginning with $\binom{n}{0} = 1$, $\binom{n}{1} = n$):

$$\begin{array}{ccccc} 1 & -3 & 6 & -10 & 15 \\ 1 & -2 & 3 & -4 & 5 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 3 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \end{array}$$

This case amounts to the following statement:

Lemma 1. *For each $k \in \mathbb{N}$, the product of any k consecutive positive integers is divisible by $k!$.*

Proof. When $n \in \mathbb{N}$, the numerator of $\binom{-n}{k}$, in reduced form, is, up to sign, the product of the k consecutive integers starting at n. □

Given two integers $a, b \in \mathbb{N}$, we may factor b as $b = rs$, where r is relatively-prime to a, and s is a product of powers of primes that divide a. This factorization is unique: r may be specified as *the largest* factor of b such that $\gcd(a, r) = 1$. We call r *the part of b prime to a*.

A direct attack on Theorem 1.2 reduces it to showing the following generalization of Lemma 1:

Theorem 4.1. *Let $d, k \in \mathbb{N}$ and let r be the part of $k!$ prime to d. Then r divides the product of the terms of each k-term arithmetic progression of integers having step d.*

Proof. For $m \in \mathbb{Z}$, let

$$f(m) = m(m + d) \cdots (m + (k - 1)d).$$

We have to show that $r | f(m)$ for each $m \in \mathbb{Z}$.

Now $\gcd(r, d) = 1$, so we may choose $e \in \mathbb{Z}$ with $de \equiv 1 \pmod{r}$. Then

$$e^k f(m) \equiv em(em + 1) \cdots (em + k - 1) \pmod{r}.$$

By Lemma 1, $k!$ divides the right-hand-side, so r divides $e^k f(m)$, so $r | f(m)$, since $\gcd(r, e) = 1$. □

BIBLIOGRAPHY

[1] J.-P. Serre, A Course of Arithmetic. Springer. New York. 1996.

THE MAGIC OF COMPLEX ANALYSIS

THOMAS P. LEAHY
National University of Ireland, Galway

1. THE HISTORY OF COMPLEX NUMBERS

Complex Analysis is all fun and games until someone loses an i.

What are complex numbers?

A complex number is a number which can be written in the form $z = a + bi$ where a and b are real numbers and i is the imaginary unit satisfying $i^2 = -1$. We call a the real part of z and b the imaginary part, written respectively as

$$\mathrm{Re}(z) = a \text{ and } \mathrm{Im}(z) = b.$$

The set of all complex numbers is denoted $\mathbb{C}$. The complex number $a + bi$ can be visually represented as a vector (a, b) on a two-dimensional plane diagram called an Argand diagram. This diagram represents the complex plane and has two axes: the real axis and the imaginary axis.

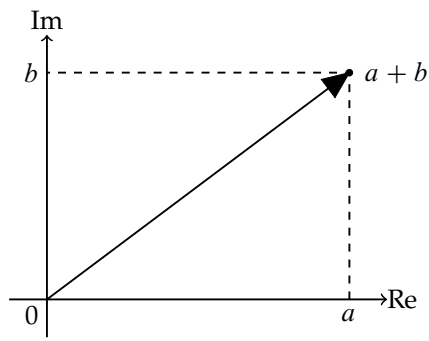


Figure 3.1: The complex plane represented on an Argand diagram.

Where did it all come from?

The very first mention of the notion of complex numbers goes back almost 2000 years. In the first century, Heron of Alexandria was trying to calculate the volume of a pyramidal frustum. To solve it he had to find the square root of a negative number, and operation he deemed impossible, so he gave up. In the 1500s, some speculation about the square roots of negative numbers returned when formulas for solving 3rd and 4th degree polynomial equations were discovered. Mathematicians found that when using these formulae to

find real solutions of certain polynomial equations some intermediate work with negative square roots would be required. At the end of the calculation these negative square roots would cancel leaving a correct real solution. This problem led at least one mathematician to publish solutions to certain equations without any clue as to how he actually found them! An example of this phenomenon, studied by the Italian mathematician Rafael Bombelli, can be seen with the equation

$$x^3 - 15x - 4 = 0.$$

Employing the cubic formula gives a solution

$$x = \left(2 + \sqrt{-1}\right) + \left(2 - \sqrt{-1}\right)$$

which formally involves imaginary terms. It simplifies to the valid real solution $x = 4$.

Another Italian mathematician Gerolamo Cardano worked with complex numbers around 1545. He attempted to find two numbers a and b satisfying

$$a + b = 10 \text{ and } ab = 40.$$

The solution to this problem is given by

$$a = 5 + \sqrt{-15} \text{ and } b = 5 - \sqrt{-15}$$

which we can easily find using the quadratic formula. These roots are certainly not real. It was becoming increasingly clear that imaginary numbers could not be avoided.

The term "imaginary" itself was coined by René Desartes (not an Italian mathematician, but French!) in the 1630s to reflect his observation that "for every equation of degree n, we can imagine n roots which do not correspond to any real quantity". This is an informal description of a key theorem in mathematics, the Fundamental Theorem of Algebra, which more precisely states that every n^{th} order polynomial with real coefficients has exactly n roots in $\mathbb{C}$.

Complex numbers gradually gained acceptance. John Wallis developed their geometric representation in 1685. The most rigorous development (and in a sense justification) was given by the Irish mathematician Sir William Rowan Hamilton. He formally defined complex numbers as pairs of numbers (a, b) with associated operations

$$(a, b) + (c, d) = (a + b, c + d) \text{ and } (a, b) \cdot (c, d) = (ac - bd, ac + bd).$$

This purely algebraic construction sidesteps the philosophical issue of the mysterious "number" i, which in Hamilton's system is associated to the pair $(0, 1)$. Many years later Hamilton extended this method of construction to a four dimensional analogue of complex numbers, quaternions.

The Birth of Modern Day Complex Analysis

Complex analysis is the branch of mathematical analysis that examines functions of complex numbers. Many complex functions are simple analogues of real functions, for instance

$$f(z) = z^2 \text{ and } g(z) = \frac{1}{z},$$

with the only difference being that we allow z to be complex. A number of prominent and well known mathematicians have contributed to complex analysis, including Euler, Gauss, Riemann and, of course, Cauchy.

There are several fundamental ideas in complex analysis. The derivative of a complex function is defined in a similar way to the real case,

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}.$$

However it turns out that it is much harder for a complex function to be differentiable than a real function. The reason is that the limit above must be the same no matter which way h approaches 0. If h is real then it can only approach 0 from two directions: from the negative direction or the positive direction. However when h is complex it can approach 0 (the origin of the complex plane in figure 3.1) from uncountably many directions!

This idea of it being difficult for a complex function to be differentiable is captured by one of the core concepts in complex analysis: the Cauchy-Riemann equations. Take some complex function $f(z)$. The number z is complex and so can be written as $z = x + yi$ for x and y real. The value of the function is also complex, and depends on x and y. We can thus write

$$f(z) = f(x + yi) = u(x, y) + v(x, y)i$$

for two functions u and v that each take in two real numbers and output a real number.

Now suppose that $f(z)$ is differentiable at a point $z_0 = x_0 + y_0 i$. Then the functions u and v *must* be first order differentiable and *must* satisfy

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \text{ and } \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{3.4}$$

If this happens we have

$$f'(z_0) = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x} i.$$

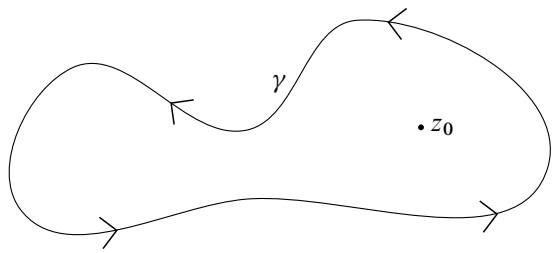
The equations in (3.4) are the Cauchy-Riemann equations. If a function satisfies them at z_0 then we say the function is *analytic* at z_0.

We said above that it is hard for a complex function to be differentiable, and the Cauchy-Riemann equations really demonstrate this. Suppose we create a complex function by randomly choosing two functions u and v – let's make them infinitely differentiable to be easy on ourselves! – and define $f(x+yi) =$

$u(x, y) + v(x, y)i$. What are the chances that this randomly chosen function will satisfy (3.4) and hence be complex differentiable? It is almost impossible! For a given u there is only one associated v that will satisfy (3.4) everywhere, while there are more than uncountably many possible choices of v to pick from initially. So even though our two component functions u and v are infinitely differentiable our complex function is almost surely not even once differentiable in the complex sense!

Now the question is, do we get anything in return for having such a stringent definition of differentiability? The answer is an emphatic yes. Let's look at an important example, an important cornerstone of complex analysis called CIF – no, not the common household cleaning product, but the Cauchy Integral formula.

Let γ be a simple closed curve in the complex plane, oriented counterclockwise. Simple means the curve doesn't intersect itself, closed that it has no endpoints, and the orientation just gives the direction we integrate over (recall for ordinary integrals we integrate left to right). So curve γ might look something like this:



Let $f(z)$ be any complex function which is analytic at all points on and inside γ. If z_0 is inside the region bordered by γ, then

$$f(z_0) = \frac{1}{2\pi i} \int_\gamma \frac{f(z)}{z - z_0} \mathrm{d}z. \tag{3.5}$$

So by just knowing the value of $f(z)$ on γ and the fact that it's analytic we can find its value anywhere inside γ. Even better, this formula extends to derivatives: for *any* n we have

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_\gamma \frac{f(z)}{(z - z_0)^{n+1}} \mathrm{d}z.$$

The Cauchy Integral Formulae have many important consequences; we list three.

- If a function f is analytic at a given point, then its derivatives of all orders are analytic there too.
- Let f be continuous on a domain D. If

  $$\int_C f(z)\mathrm{d}z = 0$$

  for every closed contour C in D, then f is analytic throughout D.

- Suppose that a function f is analytic inside and on a circle C_R, centered at z_0 and of radius R. If M_R denotes the maximum value of $|f(z)|$ on C_R, then
$$|f(z_0)| \leq \frac{n!M_R}{R^n}$$
for every $n \in \mathbb{N}$.

2. APPLICATION: REAL INTEGRALS

Integrals are an integral part of mathematics, mind the pun. Actually computing integrals is often important; however, most integrals cannot be solved by using "traditional" methods. Many of these real integrals (usually improper, meaning one or both of the limits of integration is $\pm\infty$) can be solved by using complex analytical methods. The primary tool are *residues*.

To describe residues, let's pick a complex function f that is analytic everywhere except at certain points where it "blows up". A good example is
$$f(z) = \frac{1}{z},$$
which is analytic everywhere except at $z = 0$. We say that f has a singularity at $z = 0$.

A residue is just a complex number related to the singularity. First we pick any closed curve γ_0 that encloses the singularity z_0; for our example $z_0 = 0$ and the unit circle will do as γ_0. Then the residue of f at the singularity z_0 is defined by
$$\text{Res}(f, z_0) = \frac{1}{2\pi i}\int_{\gamma_0} f(z)dz.$$

Residue Theorem

Let γ be a simple closed curve in $\mathbb{C}$, oriented couterclockwise. If a function f is analytic on and inside γ, expect possibly for a finite number of singularities at $z_1, ..., z_n$, then
$$\int_\gamma f(z)\mathrm{d}z = 2\pi i \sum_{k=1}^{n} \text{Res}(f, z_k).$$
Often this can give an easy way of evaluating real integrals which would be hard by conventional (real analytic) methods. Residues are unquestionably a useful tool for solving difficult real integrals, but sometimes the notation can be off-putting. I am going to look at evaluating these integrals in a slightly different way, using just the knowledge we already have and a little mathematical trickery. The best way to understand the method is with an example.

Example

We wish to evaluate
$$I = \int_0^\infty \frac{\mathrm{d}x}{(x^2+9)(x^2+4)}.$$

This is the same as

$$\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{(x^2+9)(x^2+4)}.$$

Now we let

$$f(z)=\frac{1}{(z^2+9)(z^2+4)}.$$

This function clearly has singularities at $\pm 2i$ and $\pm 3i$. Now, relabelling the integral, we see that we want to calculate

$$\lim_{R\to\infty}\int_{-R}^{R}\frac{\mathrm{d}z}{(z^2+9)(z^2+4)}.$$

The next step is to enclose the singularities above the real axis using the upper half circle as in figure 3.2. Let C_R be the semicircle of centre 0 and radius $R>3$ (chosen to larger than the modulus of the biggest singularity of $f(z)$).

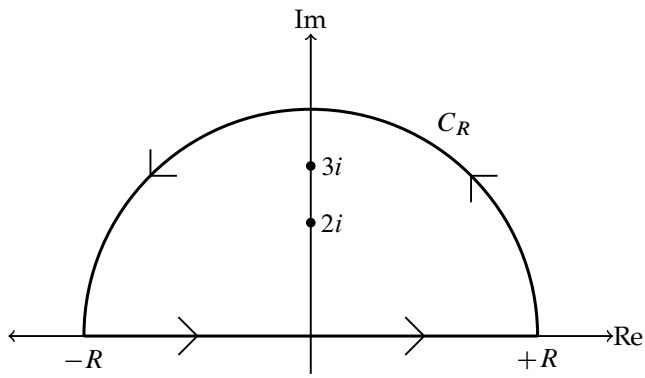


Figure 3.2: Upper Half Circle

Now consider $\gamma_R=[-R,R]\cup C_R$, oriented counter-clockwise. By the basic linearity of integration we have

$$\int_{\gamma_R}\frac{\mathrm{d}z}{(z^2+9)(z^2+4)}=\int_{-R}^{R}\frac{\mathrm{d}z}{(z^2+9)(z^2+4)}+\int_{C_R}\frac{\mathrm{d}z}{(z^2+9)(z^2+4)} \tag{3.6}$$

We can calculate the integral on the far left using CIF, and we can also show that as $R\to\infty$ the integral on the far right goes to 0. This will leave us with a value for the integral in the middle in the limit $R\to\infty$ which is exactly what we want to calculate.

We employ the Cauchy Integral formula on the left integral twice, once for each of the singularities enclosed by γ_R. How do we do this? Observe that we can factorize $f(z)$ as

$$f(z)=\frac{1}{(z-2i)(z+2i)(z-3i)(z+3i)}.$$

We are interested in calculating the integral of f along a small curve α that only contains the singularity $z = 2i$. Now inside this curve the function

$$g(z) = f(z)(z - 2i) = \frac{1}{(z + 2i)(z - 3i)(z + 3i)}$$

is analytic as all of the remaining singularites are outside the curve. Now by CIF (3.5) we have, with $z_0 = 2i$,

$$g(2i) = \frac{1}{2\pi i} \int_\alpha \frac{g(z)}{z - 2i} \mathrm{d}z,$$

But $g(z)/(z - 2i)$ is just $f(z)$, giving

$$g(2i) \times 2\pi i = \int_\alpha f(z)\mathrm{d}z,$$

and so

$$\int_\alpha f(z)\mathrm{d}z = \left(\frac{1}{(2i + 2i)(2i - 3i)(2i + 3i)} \right) 2\pi i = \frac{\pi}{10}$$

Similarly, the integral of a small curve around the singularity at $z = 3i$ is $-\pi/15$ (work it out!) and this finally gives

$$\int_{\gamma_R} \frac{\mathrm{d}z}{(z^2 + 9)(z^2 + 4)} = \frac{\pi}{10} - \frac{\pi}{15} = \frac{\pi}{30}.$$

Now we want to show that the limit as $R \to \infty$ of the right integral is zero;

$$\lim_{R\to\infty} \int_{C_R} \frac{\mathrm{d}z}{(z^2 + 9)(z^2 + 4)} = 0.$$

This is a basic exercise in bounds:

$$\begin{aligned}
\left| \int_{C_R} \frac{\mathrm{d}z}{(z^2 + 9)(z^2 + 4)} \right| &\leq \int_{C_R} \frac{|\mathrm{d}z|}{|z^2 + 9|\,|z^2 + 4|} \\
&\leq \int_{C_R} \frac{|\mathrm{d}z|}{||z^2| - 9|\,|z^2| - 4||} \\
&= \frac{1}{(R^2 - 9)(R^2 - 4)} \int_{C_R} |\mathrm{d}z| \\
&= \frac{\pi R}{(R^2 - 9)(R^2 - 4)} \to 0, \text{as } R \to \infty.
\end{aligned}$$

Subbing all this into (3.6) we get

$$\frac{\pi}{30} = \int_{-R}^{R} \frac{\mathrm{d}z}{(z^2 + 9)(z^2 + 4)} + 0, \text{as } R \to \infty$$

and hence

$$I = \int_0^\infty \frac{\mathrm{d}z}{(z^2 + 9)(z^2 + 4)} = \frac{1}{2}\frac{\pi}{30} = \frac{\pi}{60}.$$

We have solved the original integration problem.

This almost ad-hoc method for solving these types of integrals is very effective and relatively straightforward. It requires no more knowledge than we have already acquired and the simplicity of breaking the integral we want to find into integrals we can find a solution to makes this method very attractive.

The idea that a real integral can be solved by using complex analysis methods is quite amazing. Complex analysis has a vast amount of applications not only in mathematics and applied mathematics but also in physics and engineering.

3. OTHER APPLICATIONS

There are many applications of complex numbers and complex analysis. In electrical engineering, complex numbers are perfect for treating amplitudes and phases correctly as AC currents and voltages are combined. They are also used in signal processing, which has applications to telecommunications, radar (which assists the navigation of aircraft), and even biology (in the analysis of firing events from neurons in the brain). They are a fundamental aspect of the physical field of quantum mechanics. Imaginary (complex) numbers help form the descriptions of electronic states in materials which lead to applications in optics. Complex numbers and complex functions have an important role in the field of fluid mechanics. Potential flows can be dealt with very easily through functions of complex numbers.

In number theory, Riemann introduced revolutionary ideas into the subject, the chief of them being that the distribution of prime numbers is intimately connected with the zeros of the analytically extended Riemann zeta function of a complex variable. In particular, Riemann introduced the idea of applying methods of complex analysis to the study of the the prime number counting function $\pi(x)$, which gives the number of primes less than or equal to x.

4. CONCLUSION

There are many applications, some unexpected, of complex numbers and analysis in the real world. It has come from rocky origins and skeptical development, to its initial mainstream acceptance and now worldwide appreciation. Many notable mathematicians have contributed to this intriguing branch of mathematics, but it doesn't stop there. The future of complex analysis is very much active. There is a lot of current research in this area, especially in Hypercomplex analysis and Quaternionic analysis. In fact, if you fancy winning yourself one million US dollars, see if you can solve one of the millennium problems, the Riemann hypothesis.

Question

The distribution of the prime numbers among all natural numbers does not follow any regular pattern; however, Riemann observed that the frequency of

prime numbers is very closely related to the behavior of an elaborate function

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + ...$$

called the Riemann Zeta function. The Riemann hypothesis asserts that all interesting solutions of the equation

$$\zeta(s) = 0$$

occur when the real part of s is precisely one half. This has been checked for the first one and a half billion solutions. A proof that it is true for every interesting solution would shed light on many of the mysteries surrounding the distribution of prime numbers.

Did you know i^i is a real number? Check it out!

BIBLIOGRAPHY

[1] James Ward Brown and Ruel V. Churchill. *Complex Variables and Applications, Eighth Edition*. McGraw-Hill, New York, 2009. ISBN 9780073051949.

[2] Danilo Mandic and Vanessa (Su Lee) Goh. *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, West Sussex, 2009. ISBN 9780470066355.

THE FALSE POSITIVE PARADOX

MICHAEL HANRAHAN

University College Cork

The false positive paradox is a counter-intuitive statistical result whereby a positive test result is more likely to be a false positive than a true positive. This occurs when the incidence of the thing you are testing for is lower than the false positive rate of the test. This phenomenon can cause confusion when dealing with medical test results for very rare conditions. The result of it is that when you get a positive result for a rare condition the actually probability that you do have the condition is still quite small even for very accurate tests.

There are screening tests for many medical conditions and diseases that doctors use to diagnose people who may be at risk. For very rare diseases the vast majority of people screened will get a negative result and generally the risk of getting a false negative result is low. However when someone gets a positive result it does not necessarily mean that the patient actually does have the disease or condition. This causes a lot of anxiety, stress and additional financial costs for further testing. On top of this, unnecessary additional testing can be painful and lead to complications. For instance, a prostate biopsy conducted after a positive PSA test can have complications such as infection, bleeding into the urethra or bladder, inability to urinate, bleeding into the rectum, or allergic reactions to the anaesthetic.

1. BAYES' THEOREM

In order to calculate the probability of actually having a true positive result when the test is positive statisticians use Bayes' theorem. To use Bayes' theorem in relation to medical screening the following information is needed:

- The prevalence of the condition/ disease in the population being tested.
- The sensitivity (actual positives correctly identified) and specificity (actual negatives correctly identified) of the screening test.

Mathematically, the theorem is expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|-A)P(-A)}$$

where the event A signifies having the condition, the event B signifies having a positive test result, and hence

- $P(A|B)$ represents the chance of having the condition given a positive test result.
- $P(B|A)$ represents the chance of a true positive; i.e., the positive test result being correct.

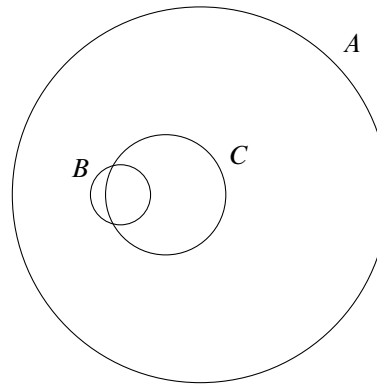


Figure 4.1: A represents the whole population to be tested, B represents the number of people who have the condition and C represents the people who get a positive test result. So it's clear that most of the people who do have the condition do get a positive result, but so do a lot of healthy people. The false positive paradox arises because C is at least twice as big as $B \bigcap C$.

- $P(A)$ represents the incidence of the condition in the population,
- $P(-A)$ represents the chance of not having the condition (i.e. 1 – P(A))
- $P(B|-A)$ represents the chance of a false positive; i.e., getting a positive test result but not actually have the condition.

In words the theorem says

$$\text{chance that a positive result is correct} = \frac{\text{number of true positives}}{\text{number of total positives}}.$$

When using the theorem it can be helpful to write the variables in table format.

	Has condition, $P(A)$	Doesn't have condition, $P(-A)$
Positive Test Result $P(B)$	True Positive, $P(B\|A)$ (test sensitivity)	False Positive, $P(B\|-A)$
Negative Test Result $P(-B)$	False Negative	True Negative (test specificity)

The false positive paradox arises because medical screening tests are never 100% specific (a screening that was specific would never give false positive) and for rare diseases a small percentage of a large number of people (number of false positives) can be larger than a high percentage of a small number of people (number of true positives). See the visualization in figure 4.1.

We now look at some real-life applications of Bayes' theorem to common screening tests.

2. FIRST EXAMPLE: BREAST CANCER

The prevalence of breast cancer in Ireland is approximately 0.2% in women aged over 40. The screening mammography has a sensitivity of 69% and a specificity of 94%. This gives rise to following the table.

	Has Breast Cancer (0.2%)	No Breast Cancer (99.8%)
Positive Test Result	69%	6%
Negative Test Result	31%	94%

We then get

$$\text{True Positive} = 69\% * 0.2\% = 0.00138,$$
$$\text{False Positive} = 6\% * 99.8\% = 0.05988.$$

We then have that the probability of having Breast Cancer after receiving a positive test result is

$$\frac{\text{true positives}}{\text{number of all positives}} = 2.25\%.$$

This is a shocking result and is the source of much controversy as to whether mammograms are even worth doing. Approximately 977 people out of 1000 people who have had a positive mammogram result do not actually have cancer. About 10% of those people will be sent for additional testing. 14% of that 10% will be sent for a biopsy, and of this group 65% will still not have breast cancer owing to the fact that the biopsy test has its own false positive rate.

Another result of counter-intuitive probability states that over a decade of annual mammogram screening about 46% of women will receive a positive result, the vast majority of which will be false positives.

We have now clearly shown some disadvantages of mammogram screening in terms of unnecessary costs and anxiety for patients. The question that remains is why do we tolerate such high false positive rates? Surely this test can be refined to give fewer false positives?

3. ROC CURVES

The reason why the test performs to these standards can be explained by the Receiver Operator Characteristic (ROC) curve of the test. The curve demonstrates that there must be a "trade-off" between sensitivity and specificity. This arises because, in general, there is no perfect marker for a disease. The levels of the marker found in any individual with or without the disease will follow a general pattern of distribution. The problem is that overlapping of the two distribution curves of healthy people and people with the disease occurs so that at a certain concentration of the marker it is difficult to tell whether or not the person has the disease.

And example distribution is illustrated in figure 4.2. The regions are:

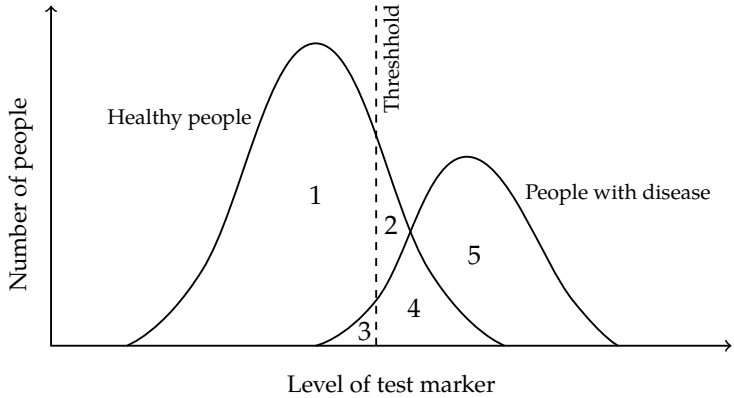


Figure 4.2: Distribution curves for a particular disease.

1,3. Healthy people who receive true negative test results.

2,4. Healthy people who receive false positive test results.

3. People with the disease that receive false negative test results.

4,5. People with the disease that receive true positive test results.

The threshold is an arbitrary cut-off point designated by the creator of the test above which any level of marker detected will yield a positive test result. The ideal threshold point maximises the number of true positives and minimises the number of false positives. However in some instances it may be more useful to have a high specificity.

In order to visualize the effect of moving the threshold point (i.e., the trade-off between sensitivity and specificity) we use the ROC curve (figure 4.3).

4. SECOND EXAMPLE: HIV

Human immunodeficiency virus (HIV) is a virus that infects the human immune system and leaves the infected person susceptible to life-threatening opportunistic infections and cancers. It can cause acquired immunodeficiency syndrome (AIDS).

The prevalence rate of HIV in adults (aged 15 to 49) in Ireland is 0.2% (2009 estimate from UNICEF). An oral home-use HIV test can be used to detect whether or not HIV antibodies are present in someone who is at risk of being infected with the HIV virus. For this test, 99.9% of negative results are true negatives (i.e., the person does not have HIV antibodies) while 91.7% of positive results are true positives. We put these numbers in table form:

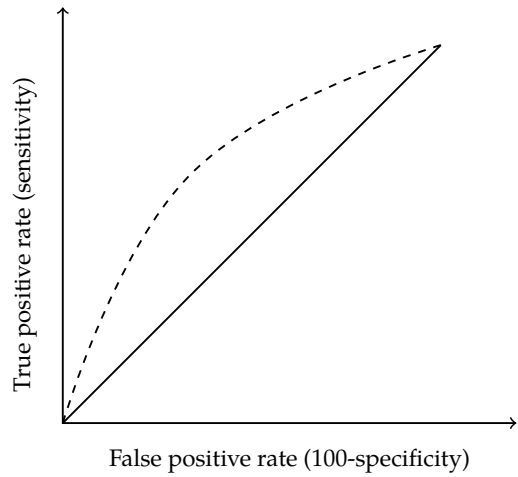


Figure 4.3: ROC curve: the more sensitive we want the test to be, the less specific it will be.

	HIV Positive (0.2%)	HIV Negative (99.8%)
Positive test result	91.70%	0.10%
Negative test result	8.30%	99.90%

What is the likelihood that you have HIV if you get a positive result? As before we calculate the numbers:

$$\text{true positive} = 91.7\% * 0.2\% = 0.001834,$$
$$\text{false positive} = 0.1\% * 99.8\% = 0.000998.$$

We then get that the probability of being HIV positive is

$$\frac{\text{true positives}}{\text{all positives}} = 0.001834/0.002832 = 64.75\%$$

Therefore, there is only a 64.75% chance that you truly are infected with the HIV virus if you get a positive result. In other words, roughly 35 people out of 100 who get a positive result using this test will actually not be infected with HIV. This is a counter-intuitive result as the test originally appeared to have a 91.7% accuracy rate. If someone does get a positive test result it cannot be inferred that they do have HIV and further testing is required. However it is easy to see how this could cause a lot of unnecessary anxiety and stress on a person.

Clearly this home test was designed to give a true negative result to people who do not have HIV. Using the same method as above it's found that a negative test result means there is a 99.99% chance that you don't have HIV.

So this wasn't a true example of the false positive paradox but it's still a very interesting example and highlights the importance of not over-reacting to a positive test result.

CONSTRUCTING THE REAL NUMBERS

DARON ANDERSON

Trinity College Dublin

Our goal is simple: to show that the set of real numbers, which we use every day and which seem self-evident from the properties of observable space, can be defined in a purely abstract way that is free from self-contradiction.

The real numbers can be intuitively understood in terms of continuous amounts, or as positions on the number line. But if we try to build a definition around this concept, removing all references to an external world, it is hard to see what are we left with. What is an amount if not an amount *of something*? The amounts on their own, considered without the associated operations of addition, multiplication, order, and absolute value, are of little interest. Indeed, in this form, the only notable property is the set's cardinality, a measure of how *large* it is.

Thinking of the set $\mathbb{R}$ like this, with no way of comparing or manipulating the elements, gets us nowhere. The characteristics of the real numbers that we are interested in - those we would like to prove make sense - include the manner in which they interact under the rules of arithmetic, how these rules interact with each other, and the specific way in which we have the set inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$. These well known simpler subsets of the reals can be defined in a purely symbolic manner. We will do this, if only to illustrate the comparative ease with which it can be done.

The natural numbers can be defined rigorously as elements of the form $1 + 1 + \ldots + 1$, with addition defined in the obvious manner. We define multiplication by setting $1 \cdot 1 = 1$, and by imposing the distributive law: that $a \cdot (b + c) = a \cdot b + a \cdot c$, for every $a, b, c \in \mathbb{N}$. This extends to $\mathbb{Z}$ simply enough by introducing one new element -1 such that $1 + (-1) = 0$ and again asserting that multiplication distributes over addition.

We then construct the rational numbers $\mathbb{Q}$ as elements of the form $\frac{a}{b}$ where a and b are integers and b is nonzero. We then *define*

$$\begin{aligned} \frac{a}{b} = \frac{c}{d} &\text{ to mean } ad = bc, \\ \frac{a}{b} + \frac{c}{d} &\text{ to be equal to } \frac{ad + cb}{bd}, \\ \frac{a}{b} \cdot \frac{c}{d} &\text{ to be equal to } \frac{ac}{bd}, \end{aligned}$$

and we're done.

Attempts to extend this technique to $\mathbb{R}$ provoke some immediate questions. For example, what exactly does it mean to say that $2.71828\ldots$ has an infinite number of decimal places? This question is especially difficult since $2.71828\ldots$ is not so much a mathematical object as a method of representing one.

We could conceivably define the reals in an algorithmic way, as maps of the form $x : \mathbb{N} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ which assign to each n what we understand to be the n^{th} decimal digit of x, along with rules for adding and

multiplying such maps. However, unlike in the case of rational numbers, these algorithms must be infinite. And it is difficult to define the result of an algorithm which never terminates. We can sidestep this problem by giving, instead of a rule for calculating the entire decimal expansion of $x + y$, say, a recursive procedure which will produce any chosen $(x + y)(n)$ after a finite number of iterations. This is a valid yet unappealing definition of the real numbers.*

Overall, there are three problems with this approach. The first is that it places undue importance on the decimal representation of numbers: mathematically speaking, there is nothing special about this way of writing real numbers. We could just as easily write the number in binary, or indeed any integer or non integer base, or use continued fractions. The second problem is that this system does not capture the fact that the decimal expansion of a real number is not necessarily unique. For instance

$$1 = 1 \cdot 1 = \frac{3}{3} \cdot 1 = 3 \cdot \frac{1}{3} = 3(3.333\ldots) = 0.999\ldots$$

and therefore $1 = 0.999\ldots$ Finally, we want to be able to prove things about the real numbers and use them to study other concepts without always having to refer back to the nitty-gritty of how addition and multiplication work. We want to be able to say *let x be a real number* without worrying about how to write x down. In order to do this, we would most likely have to prove that this construction of $\mathbb{R}$ is equivalent to some other more natural one. This then renders the original construction redundant.

1. FIRST CONSTRUCTION: AS A COMPLETION

What more natural constructions exist? The first one here relies on the fact that for each real number, there are rational numbers arbitrarily close to it. We must make rigorous the concept of estimating reals by rationals.

The first step is to present the rational numbers differently. For each $q \in \mathbb{Q}$ let $\tilde{q}$ denote the rational sequence $(q, q, q, \ldots)$. We can define addition and multiplication pointwise on the space of these sequences. This gives us a structure which behaves identically to the rational numbers.

What happens if we allow non constant sequences? Perhaps if we allow all sequences we get a collection of objects that behaves like the real numbers? We can certainly allow non constant sequences, but it turns out the space of all sequences is simply too big to be $\mathbb{R}$. For instance there is only one element of $\mathbb{R}$ that cannot be inverted - namely 0 - while there are many sequences other than $\tilde{0}$ which are non-invertible. For instance, there is nothing which when multiplied by $(0, 1, 1, \ldots)$ yields $\tilde{1}$. But we want $\mathbb{R}$ to contain only one element 0 with this non-invertibility property.

*[Editor's note]: This method of construction is also in a certain sense impossible. Just as the set of all numbers with finite decimal expansion is a countable set, so is the set of all "finite algorithms"; that is, algorithms that can be defined using a finite number of instructions. As the set of real numbers is uncountable it follows that almost all real numbers cannot be produced using a finite algorithm. These numbers are termed uncomputable. For a popular account of this issue see the second chapter of Roger Penrose's book *The Emperor's New Mind*.

However, there is a special category of sequences which suits our purpose. We allow exactly those sequences which tend towards what we would naively regard as a real number. These are called Cauchy sequences. For example $(3, 3.1, 3.14, 3.141, 3.1415, 3.14159, \ldots)$ is allowed. The terms in this sequence get arbitrarily close to each other, and we would intuitively expect this means the sequence converges. It is true that over the real numbers, this sequence converges. But since we have not defined the real numbers yet, such a statement is so far meaningless.

Clearly more than one sequence can approximate the same element of $\mathbb{R}$ and so we must consider some of the allowed sequences to be *the same*. We would like to allow $(x_m)_{m=1}^{\infty}$ and $(y_m)_{m=1}^{\infty}$ to be equivalent if

$$\lim_{m\to\infty} x_m = \lim_{m\to\infty} y_m,$$

but since these limits may not be defined the best we can hope for is that

$$\lim_{m\to\infty} (x_m - y_m) = 0.$$

This makes sense since every $x_m - y_m$ is a rational number.

If a number x is rational, we can represent it in this space by $\tilde{x}$. If we want the number to be irrational, we can use the sequence $(x_m)_{m=1}^{\infty}$ which has x_n as the decimal expansion of x truncated to n decimal places. This is clearly a rational sequence.

In order to suggest the rules of arithmetic on these Cauchy sequences work as intended, we will assume that $\mathbb{R}$ is already defined and that, rather than acting as a definition, this construction simply mimics it. Now x is represented by $(x_m)_{m=1}^{\infty}$ with limit x, and y is represented by $(y_m)_{m=1}^{\infty}$ and with limit y. Then $x + y$ is represented by sequence $(x_m + y_m)_{m=1}^{\infty}$ as

$$\lim_{m\to\infty} (x_m + y_m) = \lim_{m\to\infty} x_m + \lim_{m\to\infty} y_m = x + y.$$

The rule for multiplication is defined similarly. We finally need to define the ordering on the reals. We define $x \leq y$ to mean that there exists $M \in \mathbb{N}$ such that

$$x_m - y_m \leq 0 \text{ for all } m \geq M,$$

which allows us to order these sequences.

What we have done here is encoded the information required to represent a real number in a sequence, and then used the sequences themselves as our basic objects. Thus we have translated the problem about an infinite decimal expansion into one about infinite sequences, which are already well understood. This approach can be generalized to form what is called the completion of a metric space. You might hear the real numbers described as the completion of the rationals. This abbreviates the entire construction above.

2. SECOND CONSTRUCTION: DEDEKIND CUTS

While our previous construction was topological, the second involves more set theory. Dedekind cuts are named after their inventor, illustrating the significance of the individual contribution to this general approach.

Dedekind noticed that a real number can be completely specified by giving two sets: the set A of numbers which are less than or equal to it, and the set B of those which are strictly larger than it. Furthermore, he noticed that nothing is lost if A and B only contain the relevant rationals. He then flipped this realization on its head and defined a real number as a pair of sets of rational numbers (A, B) with the following properties:

1. A and B are disjoint.
2. Between them, A and B exclude at most one rational number.
3. B does not have a minimum element.
4. If A or B contains q_1 and q_2 and $q_1 < q < q_2$ then it contains q as well.

What should always be remembered is that when defining a Dedekind cut we never assume there is some point lying *between* A and B on the rational number line. We use the cut as an object in its own right.

The representation of an element $q \in \mathbb{Q}$ as the Dedekind cut (A, B) is particularly simple: we set

$$A = \{x \in \mathbb{Q} : x \leq q\} \text{ and } B = \{x \in \mathbb{Q} : x > q\}.$$

Thus if A has a maximum, the cut (A, B) represents a rational number. Let's try something harder. How about $\sqrt{3}$? We can set

$$A = \{x \in \mathbb{Q} : x \leq 0 \vee x^2 \leq 3\} \text{ and } B = \{x \in \mathbb{Q} : x \geqslant 0 \wedge x^2 > 3\}.$$

where $\vee$ denotes logical OR and $\wedge$ denotes logical AND. Here there is no confusion over which representations are equivalent. We simply say that $(A, B) = (C, D)$ if and only if $A = C$ and $B = D$ as sets. Then we define the operations of addition and multiplication as follows: $(A_1, B_1) + (A_2, B_2) = (A_{1+2}, B_{1+2})$ for

$$\begin{aligned} A_{1+2} &= \{a_1 + a_2 : a_1 \in A_1 \wedge a_2 \in A_2\} \text{ and} \\ B_{1+2} &= \{b_1 + b_2 : b_1 \in B_1 \wedge b_2 \in B_2\}. \end{aligned}$$

For (A, B) rational, we define $-(A, B)$ in the obvious way. Otherwise we set

$$-(A, B) = (-B, -A) \text{ for } -A = \{-a : a \in A\}, \text{ etc.}$$

This permits subtraction.

For ordering we denote by $(A_1, B_1) \leq (A_2, B_2)$ that $A_1 \subseteq A_2$. This allows us to define the sign of a cut,

$$\operatorname{sgn}(A, B) = \begin{cases} 1 & \text{if } 0 \leq (A, B) \wedge (A, B) \neq 0, \\ -1 & \text{if } (A, B) \leq 0 \wedge (A, B) \neq 0, \\ 0 & \text{if } (A, B) = 0, \end{cases}$$

which in turn allows us to talk about positive and negative numbers.

Multiplication is first defined for positive numbers through

$$(A_1, B_1) \cdot (A_2, B_2) = (A_{1\cdot 2}, B_{1\cdot 2})$$

where

$$A_{1\cdot 2} = \{a_1 \cdot a_2 : a_1 \in A_1 \wedge a_2 \in A_2 \wedge a_1, a_2 > 0\} \cup \{x \in \mathbb{Q} : x \leq 0\} \text{ and}$$
$$B_{1\cdot 2} = \{b_1 \cdot b_2 : b_1 \in B_1 \wedge b_2 \in B_2\}.$$

Multiplication is extended to all numbers by setting

$$(A_1, B_1) \cdot (A_2, B_2) = \operatorname{sgn}(A_1, B_1) \cdot \operatorname{sgn}(A_2, B_2) \cdot \big(\operatorname{sgn}(A_1, B_1)(A_1, B_1)\big) \cdot \big(\operatorname{sgn}(A_2, B_2)(A_2, B_2)\big).$$

These definitions give all the expected properties of the real numbers, such as commutativity and distribution of multiplication over addition. Thus we have built up an algebraic system which interacts with the rational numbers in the required way.

Now it is not immediately obvious that under the "continuous amounts" interpretation, that each real number can be represented by a Dedekind cut. Certainly those with nice properties such as $x^3 = 2$ can be, but do all real numbers have such clean rules for deciding exactly where they live on the continuum? Let's go back for a moment to our definition of $\mathbb{R}$ in terms of Cauchy sequences. It should be obvious that for each real x, some rationals are less than or equal to x, while some are greater. Moreover these two sets satisfy the Dedekind cut axioms. But describing them without explicit reference to x, that is the challenge. Another possible criticism of Dedekind's approach is that it doesn't say what a real number looks like, or how to write one down on a page. This is a blessing in disguise, since it leaves us free to use whatever notation we like, provided it makes logical sense in this framework.

3. THIRD CONSTRUCTION: ALGEBRAICALLY

The next construction of $\mathbb{R}$ requires some background, and is if anything more abstract than the previous. But since the following required definitions are already ubiquitous across mathematics, if you have not seen them before you will again soon.

The first is the notion of a group, which is a generalization of the rational numbers. It is a set of elements which can be "multiplied" together in a manner which satisfies many familiar properties. Common examples of groups include $\mathbb{Z}$ and $\mathbb{Q}$ under addition, and $\mathbb{Z} \setminus \{0\}$ and $\mathbb{Q} \setminus \{0\}$ under multiplication. (We must omit 0 in the last examples as the group operation is required to be invertible.) Slightly less common examples are the set of invertible square matrices of a given size, or injective maps from a set onto itself under composition.

Formally, a group is defined as a pair $(G, \cdot)$ consisting of a set G and a binary operation $\cdot$ on G fulfilling the following axioms

1. For all $a, b \in G$, $a \cdot b \in G$. (Closed under multiplication.)

2. For all $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ (Multiplication is associative.)

3. There exists $e \in G$ such that for all $a \in G$, $e \cdot a = a = a \cdot e$ (Identity element.)

4. For all $a \in G$ there exists $a^{-1} \in G$ such that $a^{-1} \cdot a = e = a \cdot a^{-1}$. (Invertibility.)

If, in addition, $a \cdot b = b \cdot a$ for every pair of elements $a, b \in G$ then G is called an Abelian group and we generally write + instead of $\cdot$.

A field $(F, +, \cdot)$ consists of a set F equipped with two binary operations + and $\cdot$, where $(F, +)$ is an abelian group with identity element denoted by 0, and $(F \setminus \{0\}, \cdot)$ is an abelian group with identity element denoted by 1, and the field elements also satisfy the distributive law $a \cdot (b + c) = a \cdot b + a \cdot c$.

A field is said to be totally ordered if it is equipped with an abstract binary relation, denoted $\leq$, such that

1. For all $a, b \in F$, $a \leq b \vee b \leq a$. (Totality.)

2. For all $a, b \in F$, $a \leq b \wedge b \leq a \Rightarrow a = b$. (Antisymmetry.)

3. For all $a, b, c \in F$, $a \leq b \wedge b \leq c \Rightarrow a \leq c$. (Transitivity.)

4. For all $a, b, c \in F$, $a \leq b \Rightarrow a + c \leq b + c$. (Order commutes with sum.)

5. For all $a, b \in F$, $0 \leq a \wedge 0 \leq b \Rightarrow 0 \leq a \cdot b$ (Order commutes with product.)

Since we can add the 1 of the field, or its negative, to itself n times, we can generate a copy of the integers living inside our ordered field. Phrased precisely, each ordered field contains a subfield homomorphic to $\mathbb{Z}$.

Finally, we define an Archimidean ordered field: an ordered field is said to be Archimidean if, for any element x of F, there is a smallest integer greater than x.

The point of all of these definitions is that it can be shown there exists a maximal Archimidean ordered field, meaning one which contains a copy of any other Archimidean ordered field. This unique field can be used as a definition of the real numbers.

This approach is based solely on abstract algebra. Most textbooks on real analysis will begin with a list of axioms for the real numbers. And if you look hard enough at the definitions above, you can recover most of these axioms. What the algebra gives us is that these axioms define an unique structure. This is not just our favourite Archimidean ordered field; it is the *only* Archimidean ordered field!

4. FOURTH CONSTRUCTION: SURREAL NUMBERS

Our final construction is called the Theory of Surreal Numbers. It differs from the previous three in two ways. First, it does not contain any undefined notion of equality. Second, this schema does not rely on us already having established the natural numbers. It simultaneously generates the naturals, integers, rationals, and reals as well as giving a rigorous definition of infinite numbers and infinitessimals.

This system deals with the field of *numbers* which are defined recursively as follows: $\langle \emptyset | \emptyset \rangle$ is a number and so is any symbol of the form $\langle X | Y \rangle$ where X and Y are sets of numbers such that if $x \in X$ and $y \in Y$ then $x \not\leq y$. This becomes more complicated when you realise we haven't yet defined what the statement $x \leq y$ means. This total ordering of the set of numbers is also defined recursively. We first let $\langle \emptyset | \emptyset \rangle \leq \langle \emptyset | \emptyset \rangle$. Then for $x = \langle X_L | X_R \rangle$ and $y = \langle Y_L | Y_R \rangle$ we write $x \leq y$ if every $x_L \in X_L$ satisfies $y \not\leq x_L$ and every $y_R \in Y_R$ satisfies $y_R \not\leq x$. But enough theory. Let's actually construct some numbers. First, the naturals.

- Let 0 abbreviate $\langle \emptyset | \emptyset \rangle$.
- Let 1 abbreviate $\langle 0 | \emptyset \rangle = \langle \langle \emptyset | \emptyset \rangle | \emptyset \rangle$.
- Let 2 abbreviate $\langle 1 | \emptyset \rangle = \langle \langle \langle \emptyset | \emptyset \rangle | \emptyset \rangle | \emptyset \rangle$.
- ...
- Let the numeral $n + 1$ abbreviate the number $\langle n | \emptyset \rangle$.

Next the negative integers.

- Let -1 abbreviate $\langle \emptyset | 0 \rangle = \langle \emptyset | \langle \emptyset | \emptyset \rangle \rangle$
- Let -2 abbreviate $\langle \emptyset | -1 \rangle = \langle \emptyset \langle \emptyset | \langle \emptyset | \emptyset \rangle \rangle \rangle$
- ...
- Let the numeral $-(n + 1)$ abbreviate the number $\langle \emptyset | -n \rangle$

Note that we really should write $\langle \{0\} | \emptyset \rangle$ instead of $\langle 0 | \emptyset \rangle$ and so on, but we will omit the curly braces for the sake of convenience.

Before we go any further, let's check these satisfy the order relations we are looking for. Returning to our definitions, we shall prove $0 \leq 1$; that is,

$$0 = \langle \emptyset | \emptyset \rangle = \langle X_L | X_R \rangle \leq \langle Y_L | Y_R \rangle = \langle 0 | \emptyset \rangle = 1.$$

Indeed, since X_L and Y_R are both empty, both properties are vacuously true.

As that was rather easy, we will try $1 \leq 2$; that is,

$$1 = \langle 0 | \emptyset \rangle = \langle X_L | X_R \rangle \leq \langle Y_L | Y_R \rangle = \langle 1 | \emptyset \rangle = 2.$$

Now Y_R is still empty, so the second property is true. For the first we need to prove that for every $x_L \in \{0\}$, $\langle 1 | \emptyset \rangle \not\leq x_L$. Since the only element of $\{0\}$ is 0, we then need to prove $0 \leq 2$. The proof of this is vacuous as for $0 \leq 1$.

Using standard induction we can establish the integers and show their order behaves as desired. Let's go further and define powers of $\frac{1}{2}$.

- Let $\frac{1}{2}$ abbreviate $\langle 0|1\rangle$
- Let $\frac{1}{4}$ abbreviate $\langle 0|\frac{1}{2}\rangle$
- ...
- Let $\frac{1}{2^{n+1}}$ abbreviate $\langle 0|\frac{1}{2^n}\rangle$

Now we need a way to add and multiply things like $\langle X|Y\rangle$. Again this is done recursively. First we set

$$0 + x = \langle \emptyset|\emptyset\rangle + \langle X_L|X_R\rangle = \langle X_L|X_R\rangle = x.$$

We then define

$$X + Y = \langle X_L|X_R\rangle + \langle Y_L|Y_R\rangle = \langle X_L + y, x + Y_L|X_R + y, x + Y_R\rangle$$

and allow subtraction through

$$-0 = 0 \text{ and } -X = -\langle X_L|X_R\rangle = \langle -X_L|-X_R\rangle.$$

We define multiplication by setting

$$X \cdot Y = \langle X_L|X_R\rangle \cdot \langle Y_L|Y_R\rangle = \langle Z_L|Z_R\rangle$$

where

$$Z_L = \{y \cdot X_L + x \cdot Y_L - X_L \cdot Y_L, y \cdot X_R + x \cdot Y_R - X_R \cdot Y_R\} \text{ and}$$
$$Z_R = \{y \cdot X_L + x \cdot Y_R - X_L \cdot Y_R, x \cdot Y_L + y \cdot X_R - X_R \cdot Y_L\}.$$

These rules of arithmetic are a little difficult to follow so, if you like, it may help to try to add or multiply some small numbers.

We can use the integers and powers of a half to construct what are called the dyadic fractions, fractions which have a power of two as their denominator. We can then define a real number in a way similar to a Dedekind cut, choosing $r = \langle X|Y\rangle$ where X is the infinite set of dyadic fractions we want to be less than or equal to r, and Y the set of dyadic fractions we want to be greater. This determines r.

Once again, not every $\langle X|Y\rangle$ symbol represents a unique *number*. What we can say is that if $\langle X_L|X_R\rangle \leq \langle Y_L|Y_R\rangle$ and $\langle Y_L|Y_R\rangle \leq \langle X_L|X_R\rangle$ then it makes sense to consider $\langle X_L|X_R\rangle$ and $\langle Y_L|Y_R\rangle$ as equivalent. Since we haven't used the notation yet, we can denote this relation by $\langle X_L|X_R\rangle = \langle Y_L|Y_R\rangle$.

Remember that we originally defined $\langle X|Y\rangle$ by induction. It is clear that if we have an infinite list of numbers which we wish to construct, then we should be able to do this through some form of induction involving the $\langle .|.\rangle$ operator. But one property of the real numbers is that they are uncountable meaning they cannot be arranged on a list. Considering this, it seems unlikely

that we will be able to build all of $\mathbb{R}$ in this manner using standard induction. For this we will require an axiom governing *transfinite induction*. This allows chain-of-effect type proofs where our dominoes are not necessarily discrete. It allows such deductions as that $\langle 1, 2, ...|\emptyset\rangle$ is a number, even though it can not be reduced to something in terms $\langle\emptyset|\emptyset\rangle$, as every integer can.

While the soundness of transfinite induction seems like a strange thing to believe, it is an equivalent assumption to about a million other things, some of which seem more intuitive, including the notorious Axiom of Choice.

The theory of Surreal Numbers does not rely on any previously constructed mathematical objects, though it does require some logical axioms to function. And in order to follow the proofs, one requires explicit knowledge of set theoretic axioms, which explains why it is not often taught to undergraduates. This is not to say that the previous three constructions do not contain set theory; rather in those cases it is hidden below the surface, so that it seems more like common sense than anything else.

5. AND ON...

The real numbers can be defined in many equivalent, but seemingly unrelated ways. They may not be an inherent property of the natural world. But the fact that there are so many different coloured roads to the same destination suggests that they live somewhere at the heart of our mathematical logic.

So once more from the top: let x be a real number...

ALL KINDS OF PI

JAMES FENNELL

University College Cork

As a reader of this article you will be familiar with the constant π. In fact, you are probably *so* familiar with π that you've perhaps become a little complacent about it. You know that π is irrational, but – unlike $\sqrt{2}$ – you're not exactly sure how you'd go about proving it. You know that, as an irrational number, π has an infinite and non-repeating decimal expansion, but will likely struggle to give anything more than the first 3 or 4 digits. But you're at least sure that your personal inability to recall these specific facts is irrelevant: of all mathematical concepts π is surely the one with the most static and objective existence, its eternal character always there in the background ready to be recalled when needed. The goal of this article is to show that, in thinking that, you are in a specific sense wrong.

In the following pages we will see that, rather than being some universal unchanging entity, the idea of π as we know it is inextricably linked with our personal notions about the geometry of space – and is hence subject to same kind of arbitrariness that those notions are. We will realize that statements like

$$\pi = 3 \text{ or } \pi = \sqrt{15} \text{ or } \pi = \frac{15}{4} \tag{6.7}$$

are not necessarily the ramblings of numerically illiterate engineers (though you should *always* double-check) but instead represent a different and more general way of looking at the π concept. Indeed, we will see that (6.7) are all legitimate values of π but that – and I say this now to avoid the impression that things will soon become a free-for-all – we reasonably cannot have

$$\pi = \sqrt{8} \text{ or } \pi = \frac{21}{4}.$$

Read on to see why!

It is natural to ask if this tinkering with such an age-old concept is anything more than mathematical trouble-making. It will actually turn out that, apart from being intriguing for its own sake, the idea of generalizing π will leave us with a charming if limited mathematical tool.

1. FROM WHERE COMES π?

To begin, we need to think about where π comes from. There are multiple ways of defining it. The first geometric way is to reason that, as a two-dimensional plane figure, the area of a circle should be proportional to the square of its radius, and set π to be the proportionality constant to give

$$\text{area of circle of radius } r = \pi r^2.$$

We can use integration theory to give precise meaning to the term area and, indeed, us as the definition itself

$$\pi := 4 \int_0^1 \sqrt{1 - x^2}\, \mathrm{d}x.$$

This last formulation is nice because it does not depend on intuitive geometric notions which, as seen frequently in the history of mathematics, can often be a barrier to mathematical rigour. The integral definition can be made purely formal. Keeping this mindset we can design another definition free of geometry: formally define the function

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!},$$

which makes sense as the infinite series converges for every $x \in \mathbb{R}$ by the ratio test. Then set π to be twice the smallest positive number x satisfying $\cos(x) = 0$. This definition is interesting because there is no *a priori* reason to believe that this equation actually has a solution and hence that the number π exists.

However when people are thinking about π it is rarely in terms of the zeros of the cosine function. The most common way of viewing the origin of π is in its relation to the circumference of a circle. We can define

$$\pi = \frac{1}{2r} \left[\text{circumference of any circle of radius } r\right]. \tag{6.8}$$

For this to make sense we need to define what we mean by "circle of radius r" and "circumference". When we do this we become aware of the certain arbitrariness in our value of π. Let us see how. By a circle centered at x of radius r we mean the set of all points a distance r from x, where by the distance from points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in $\mathbb{R}^2$ we mean

$$\|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \tag{6.9}$$

But why should we be obliged to use this specific notion of distance?

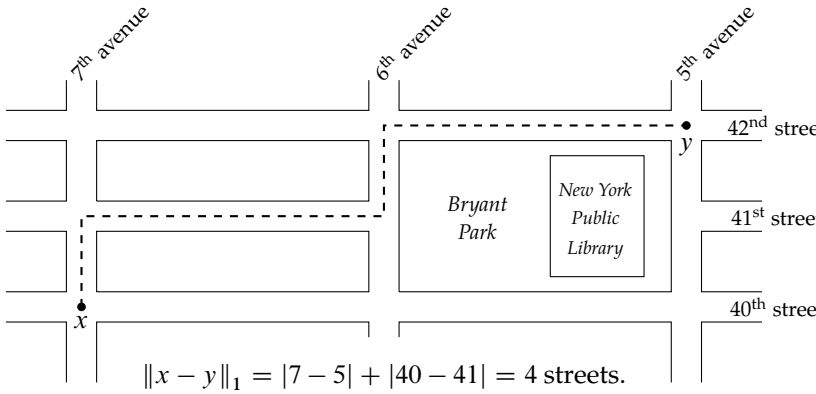
The best known alternative to the standard distance is the taxicab or Manhattan norm, where we define the distance between the two points x and y in the plane as

$$\|x - y\|_1 = |x_1 - y_1| + |x_2 - y_2|. \tag{6.10}$$

The reason for the subscript 1 will become apparent later on. Now this alternative distance has all of the usual properties that we should expect from any kind of distance function. It is always positive and it is zero if and only if x and y are the exactly same point. It satisfies the triangle inequality: if z is a third point then getting from x to y by passing through z will always take at least longer than going from x to y directly. We can write this symbolically as

$$\|x - y\|_1 \leq \|x - z\|_1 + \|z - y\|_1 .$$

Though (6.9) has the strength of the Pythagorean theorem behind it, there are scenarios in which (6.10) is more appropriate; namely if you are walking through Manhattan and can't travel "as the crow flies".



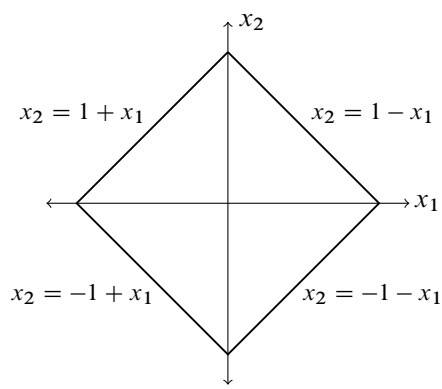
If we decide to amend our notion of distance, what will the "circles" look like? To avoid confusion when working in the more general setting, we use the term ball instead of disc, and the term boundary of the ball instead of circle. The boundary of the unit ball – the ball centered at the origin and of radius 1 – is defined analogously to the usual case as the set

$$\left\{ x \in \mathbb{R}^2 : \|x - 0\| = \|x\| = 1 \right\}.$$

If we denote the boundary of the unit ball with respect to $\|\cdot\|_1$ by B_1, then

$$B_1 = \left\{ x = (x_1, x_2) \in \mathbb{R}^2 : |x_1| + |x_2| = 1 \right\}.$$

In the upper right quadrant of the plane, $x_1 \geq 0$ and $x_2 \geq 0$, so points on the unit ball will satisfy $x_2 = 1 - x_1$. In the upper left quadrant we have $x_1 \leq 0$ and $x_2 \geq 0$, and then as $|x_1| = -x_1$ points will satisfy $x_2 = 1 + x_1$. By considering the other quadrants we can find equations for the unit ball in each case, and then see that the unit ball B_1, far from being an ordinary circle, must be a diamond:



What is the circumference of B_1? To be consistent we should measure B_1 with respect to our new distance measure $\|\cdot\|_1$. Calculating the length of B_1 is particularly simple as it is composed entirely of 4 straight lines connecting the 4 points where the boundary of the ball intersects the axes. The length of the segment in the upper right quadrant is

$$\|(1,0)-(0,1)\| = \|(1,-1)\| = |1|+|-1| = 2.$$

This is the same for all 4 lines so the circumference of B_1 is $4\times 2 = 8$. Plugging this into our definition of π in (6.8), noting that as the boundary of the unit ball b_1 as "radius" 1, we have

$$\pi_1 = \frac{1}{2}\,[\text{circumference of unit ball}] = \frac{1}{2}(8) = 4.$$

The value of π implicitly depends on how we measure distance!

2. BEYOND THE MANHATTAN NORM

The two norms we have been considering so far – the usual Euclidean norm and the Manhattan norm – are just two special cases of a general family of norms called the p-norms. The p-norm is defined for every $p \in [1,\infty)$ by

$$\|x\|_p = \left(|x_1|^p + |x_2|^p\right)^{1/p}.$$

If we plug in $p = 1$ and $p = 2$ we recover equations (6.10) and (6.9) respectively. As before, this way of measuring distance has the usual nice properties. The triangle inequality

$$\|x+y\|_p \le \|x\|_p + \|y\|_p$$

follows from the Hölder inequality, which says that if p and q are positive numbers satisfying

$$\frac{1}{p}+\frac{1}{q} = 1 \tag{6.11}$$

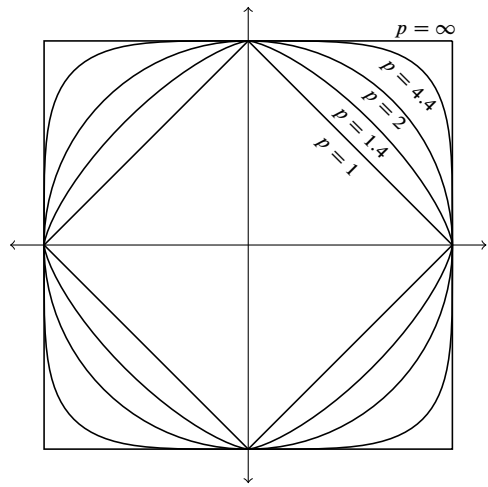
then for any $n \in \mathbb{N}$ and $x, y \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} |x_i\, y_i| \le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q}.$$

There's no need to get hung up on the specifics here, just to know that $\|\cdot\|_p$ behaves pretty much as you'd expect distance to behave. The relation (6.11) says that p and q are *Hölder conjugate*. It will reappear later! Finally, if we fix some point x and take the limit as $p \to \infty$ we are left with the supremum norm,

$$\|x\|_\infty = \max\{|x_1|, |x_2|\}.$$

The unit balls of the p-norm for a selection of values of p are drawn below. The smallest unit ball is for the $p = 1$ norm, which we have seen above. The unit balls get monotonically larger with p until we are left with the $p = \infty$ ball, which is a square.



We wish to find the π values for these norms, but calculating the length of the circumference of these balls is a far more difficult challenge then previously. We can clearly see that unless $p = 1$ or $p = \infty$ the balls aren't composed of easy to measure straight lines. This problem prompts the more general question: given a way of measuring distance in the form of a norm $\|\cdot\|$ and a continuous curve in the plane, how do we measure the length of that curve with respect to the norm?

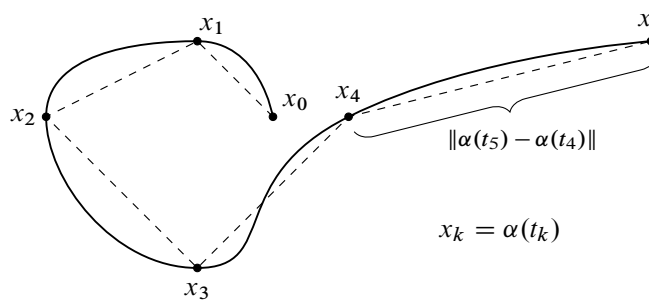
First, let us write the curve as a continuous function $\alpha : [0, 1] \to \mathbb{R}^2$. (Everything that follows actually works for curves in $\mathbb{R}^n$ but we will set $n = 2$ for simplicity.) We can approximate the curve α by fixing a partition

$$0 = t_0 \leq t_1 \leq \ldots \leq t_m = 1 \text{ of } [0, 1],$$

giving rise to a sequence of points

$$x_0 = \alpha(t_0), \ldots, x_m = \alpha(t_m)$$

on the curve α which we can join by a polygonal path in the following fashion:



Our approximation of the length of the curve is the length of the polygonal path:

$$\sum_{i=1}^{m} \|\alpha(t_i) - \alpha(t_{i-1})\| .$$

To get the actual length of the curve we then take the supremum (informally the maximum) over all possible partitions of $[0, 1]$:

$$\text{length}(\alpha) = \sup\left\{\sum_{i=1}^{m} \|\alpha(t_i) - \alpha(t_{i-1})\| : \{t_i\}_{i=0}^{m} \text{ a partition of } [0,1]\right\}$$

It is geometrically clear that this does give the correct length, but what a horrendous thing to calculate! How do even start in applying this definition of any of the unit balls above?

This problem is actually very tractable. First, the continuity of α means we don't need to take the supremum over *all* partitions of $[0, 1]$, but it is sufficient to take it over equally spaced partitions. Then

$$\text{length}(\alpha) = \sup\left\{\sum_{i=1}^{m} \left\|\alpha\left(\frac{i}{m}\right) - \alpha\left(\frac{i-1}{m}\right)\right\| : m \in \mathbb{N}\right\}. \tag{6.12}$$

This is a significant simplification.

Now let us assume that α is differentiable, with derivative $\alpha' : [0, 1] \to \mathbb{R}^2$, and fix $m \in \mathbb{N}$. We can make an appeal to the mean value theorem to say that for every integer $i \in [1, m]$ there exists

$$s_m^i \in \left[\frac{i}{m}, \frac{i-1}{m}\right]$$

such that

$$\alpha'(s_m^i) = \frac{\alpha\left(\frac{i}{m}\right) - \alpha\left(\frac{i-1}{m}\right)}{\left(\frac{i}{m}\right) - \left(\frac{i-1}{m}\right)} = m\left[\alpha\left(\frac{i}{m}\right) - \alpha\left(\frac{i-1}{m}\right)\right]. \tag{6.13}$$

This isn't strictly speaking correct as we really have to apply the mean value theorem separately to each component function of α, but the general argument isn't significantly altered. Now we can plug (6.13) into (6.12), giving

$$\text{length}(\alpha) = \sup\left\{\sum_{i=1}^{m} \frac{1}{m} \left\|\alpha'(s_m^i)\right\| : m \in \mathbb{N}\right\},$$

The quantity

$$\sum_{i=1}^{m} \frac{1}{m} \left\|\alpha'(s_m^i)\right\|$$

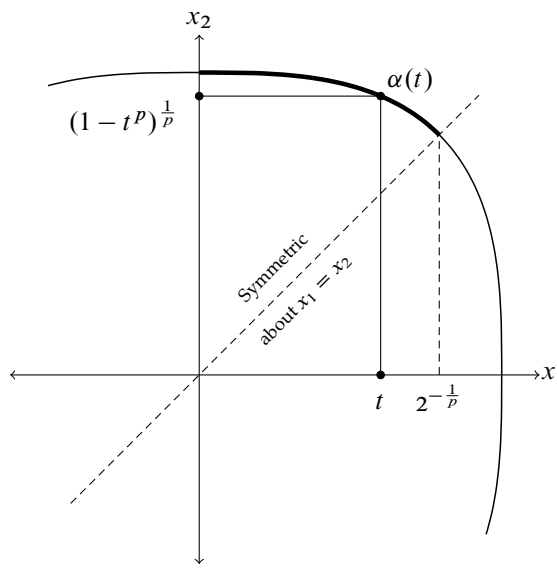
is just a Riemann sum for the integral of the function $t \mapsto \|\alpha'(t)\|$ which exists as the function is continuous. Thus taking the supremum yields

$$\text{length}(\alpha) = \int_0^1 \left\|\alpha'(t)\right\| \, \mathrm{d}t.$$

This is a much simpler expression! Readers who have studied the differential geometry of curves and surfaces may recognize it; in that subject it is called the arc-length formula. It is interesting to see that the formula does not rely on the Euclidean distance function used in classical differential geometry but is a general formula that holds in all real normed vector spaces of finite dimension.

3. π IN THE P-NORMS

We are now ready to calculate π_p, the value of π when we use the p-norm. To do this we first need to parameterize B_p, the boundary of the unit ball of the p-norm; that is, find a continuous function $\alpha : [0, 1] \to \mathbb{R}^2$ whose image is B_p. We can take a shortcut by using the multiple symmetries of the unit ball; this way we only need to parameterize one-eighth of B_p. From the diagram,



we see that

$$\alpha(t) = \left(t, (1-t^p)^{1/p}\right),\ t \in \left[0, 2^{-1/p}\right]$$

does the trick. (You can check that $\|\alpha(t)\|_p = 1$ and hence that $\alpha(t) \in B_p$ for every t.) We have

$$\alpha'(t) = \left(1, -\frac{1}{p}(1-t^p)^{1/p-1}\, pt^{p-1}\right) = \left(1, -(1-t^p)^{1/p-1}\, t^{p-1}\right)$$

and hence

$$\left\|\alpha'(t)\right\|_p = \left[1 + \left(\frac{t^p}{1-t^p}\right)^{p-1}\right]^{1/p}.$$

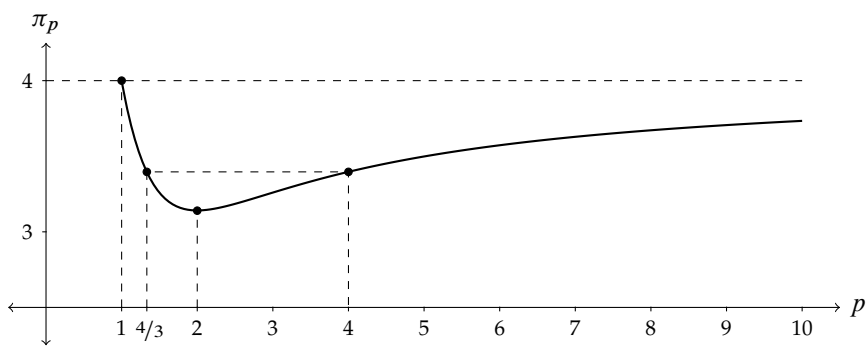
Then

$$\begin{aligned}\pi_p &= \frac{1}{2}\,[\text{circumference of unit ball}]\\ &= \frac{1}{2}\left[8 \times \text{length of } \alpha\right]\\ &= 4\int_0^{2^{-1/p}} \left[1 + \left(\frac{t^p}{1-t^p}\right)^{p-1}\right]^{1/p} \mathrm{d}t.\end{aligned}$$

The integral substitution $u = 2t^p$ simplifies the integral somewhat, giving

$$\pi_p = \frac{2}{p}\int_0^1 \left[u^{1-p} + (1-u)^{1-p}\right]^{1/p} \mathrm{d}u. \tag{6.14}$$

This expression is complicated, to say the least! Rather than trying to investigate it analytically, the easiest thing to do is plot it. This involves numerically integrating the expression (6.14) in your favourite mathematical software package for a selection of values of p. The result looks like this:



There is some notable behaviour. First, π_p appears to be restricted to the interval [3,4]. It has a maximum at $p = 1$ where $\pi_1 = 4$ (this we've already calculated) and seems to decrease monotonically to $p = 2$ where it has a global minimum of

$$\pi_2 = \pi = 3.14159\ldots$$

After $p = 2$ it increases monotonically. Calculating π_∞ in the same way we calculated π_1 (recall the unit ball for π_∞ is an easily measured square) we find $\pi_\infty = 4$. It would appear that the value of π_p asymptotically tends to this value as $p \to \infty$.

Perhaps the most interesting behavior relates to the values of p for which π_p is the same. In the diagram we that $p = 4$ and $p = 4/3$ provide an example. From

$$\frac{1}{4} + \frac{1}{4/3} = \frac{1}{4} + \frac{3}{4} = 1$$

we see that these p values are Hölder conjugate of each other, following the definition in (6.11). This turns out to be a general fact: if $p \neq q$ then we have

$\pi_p = \pi_q$ if and only if p and q are Hölder conjugate! Observing that 2 is the only number Hölder conjugate to itself,

$$\frac{1}{p} + \frac{1}{p} = 1 \implies p = 2,$$

we see that $p = 2$ is the only value for which π_p is unique.

One way to prove this Hölder conjugate fact would be manipulate the representation of π_p in (6.14). As might be expected, such an endeavor is difficult because of the complexity of the integral representation. However, this approach is unnecessary as it turns out that this fact is actually just a special case of a deeper equivalence of π values. To develop such deeper results we must stop restricting ourselves to the p-norms and ask: for *any* 2-dimensional real normed vector space, what can we say about its π value?

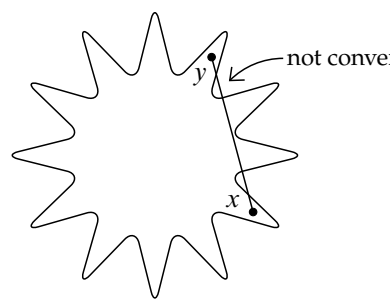
4. π IN GENERAL

So far we have focused on generalized π for given specific distance functions, namely the p-norms. As modern mathematicians we should like to study the π concept in the abstract, and see if we determine general characteristics of it independent of specific norms we look at.

In the previous section we saw that π_p was bounded between 3.14159... and 4. Is it bounded generally? The answer turns out to be yes: if we want the very intuitive triangle inequality to hold we *must* have

$$3 \leq \pi \leq 4. \tag{6.15}$$

Why? The answer is geometric. The triangle inequality forces the unit ball to be convex: if the interior of the unit ball contains points x and y it must contain every point on the straight line joining them. This means in particular that the boundary of the unit ball can't be made long by wiggling it about; for instance, it can't do this:

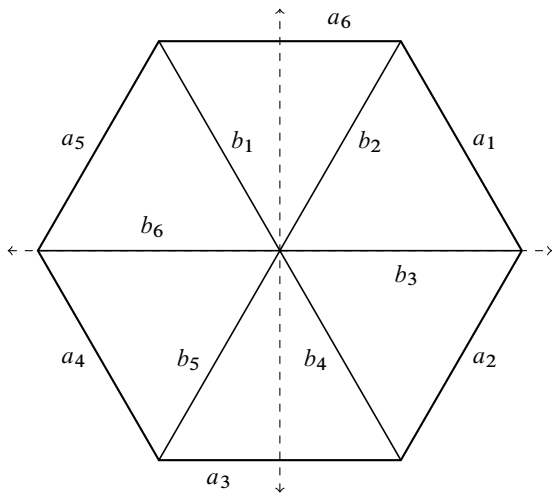


Convexity sets an upper bound on how long the boundary can be.

The existence of a lower bound is easier to justify. As the unit ball has radius of 1, any two points on the boundary of the unit ball exactly opposite each other will be a distance 2 apart. Going between these two points by travelling along the boundary of the ball will take at least as long as travelling along the

straight line joining them, and hence the total circumference of the unit ball will be at least $2 \times 2 = 4$. The π value must be at least $4/2 = 2$.

We have already seen that the upper bound in (6.15) is attained: in the very first section we found a norm, the Manhattan norm, for which $\pi = 4$. Is the lower bound attained? The answer is yes, and proving this is even easier, once you know how to draw the ball.



by translation invariance

$$L(a_i) = L(b_i) = 1$$

$$\Downarrow$$

$$\pi = \frac{1}{2}\sum_{i=1}^{6} L(a_i) = 3$$

The last general fact we will present here demonstrates the deeper reason that $\pi_p = \pi_q$ if p and q are Hölder conjugate. If we take any normed vector space X – informally $\mathbb{R}^n$ together with some norm – we can form a *dual space* consisting of the set of all linear maps from X to $\mathbb{R}$. Linear means that if $x, y \in X$ and $\alpha, \beta \in \mathbb{R}$ then

$$\phi(\alpha x + \beta y) = \alpha\phi(x) + \beta\phi(y).$$

The set of such maps, the dual space, is again a normed vector space. It has the same dimension of the original space, and comes automatically equipped with a norm induced from the original norm. In our context, this means that if we take any norm on the plane $\mathbb{R}^2$ the dual space will be $\mathbb{R}^2$ again, but with a different norm. We have the following fact:

the π value of a space is exactly the same as the π value of its dual space.

This is related to the previous section by the following theorem: the dual space of $\mathbb{R}^2$ with the p-norm is precisely $\mathbb{R}^2$ with the q-norm where q is the Hölder conjugate of p. This explains why numbers which are Hölder conjugates of each other have the same π_p value.

5. IS GENERALIZED π GOOD FOR ANYTHING?

Throughout this article we have seen that the value of π depends on our intuitive idea of distance, and hence that by looking at different ways of measuring distance we get a different value of π. In the article introduction I claimed

that this playing around with π has some actual mathematical application and wasn't merely poking holes in one of mathematics' most prized concepts for nothing. In fact, I began studying the generalized π concept not for its own value but out of a desire to find a more geometric way of solving a particular kind of mathematical problem: classifying normed vector spaces.

It is hard to describe this briefly in an accessible way, so I will just give a basic sketch. We have already seen the idea of a norm. An isometry between two normed vector spaces X and Y is a continuous bijective linear map $T : X \to Y$ which preserves distance: for all $x \in X$,

$$\|x\|_X = \|T(x)\|_Y .$$

If such an isometry exists the two spaces X and Y are said to be isometric and are considered, from the point of view of functional analysis, to be identical. My project started by examining $\mathbb{R}^n$ with the p-norms, and seeing which of these spaces are isometric. Trying to directly prove two different spaces are not isometric turns out to be quite hard, even when it is geometrically obvious that they aren't. So why is π relevant?

Consider B_X, the boundary of the unit ball of the normed vector space X. If we apply our isometry to B_X we must get some subset of B_Y as the distance value of 1 is preserved for each point. This subset must be the whole of B_Y as T is surjective (onto): any point in B_Y must be the image of some point in X, and this point in X must have distance 1 as T is an isometry. Thus, in symbols,

$$T(B_X) = B_Y .$$

Now if we take any parameterization $\alpha : [0, 1] \to X$ of B_X then the function

$$T \circ \alpha : [0, 1] \to Y$$

is a parameterization of B_Y. Then

$$\begin{aligned} \text{circumference of } B_Y &= \int_0^1 \left\| (T \circ \alpha)'(t)) \right\|_Y \mathrm{d}t \\ &= \int_0^1 \left\| T(\alpha'(t)) \right\|_Y \mathrm{d}t \\ &= \int_0^1 \left\| \alpha'(t) \right\|_X \mathrm{d}t \\ &= \text{circumference of } B_X . \end{aligned}$$

The circumferences are the same, so the π values must be the same. The number π is thus an *isometric invariant*: if two spaces are isometric they must have the same π value. An easy route to proving two spaces are not isometric is then to show that their π values are different. In our case this gives: if $p \neq q$ and p and q are not Hölder conjugate then $\mathbb{R}^2$ with the p-norm is *not* isometric to $\mathbb{R}^2$ with the q-norm.

6. ACKNOWLEDGMENTS

This article is an offshoot of my final year mathematics project at University College Cork. I would like to thank my project supervisor Dr. Stephen Wills for the vast amount of attention he gave my work, and also for his incredibly rigorous approach to mathematics which I have aspired to during my time at UCC. My thanks also to the two secondary school students who endured me over coffee in UCC one morning, and who made me realize that you don't need to know what a normed vector space is to be intrigued by the idea of π having different values.

LARGE DEVIATION THEORY: ESTIMATING THE PROBABILITIES OF RARE EVENTS

BRENDAN WILLIAMSON

Dublin City University

In statistics and probability we are often interested in the "long run" behaviour of certain objects, be they estimators of certain quantities, random walks, stochastic models or any other types of random mathematical objects. In the most elementary of these cases, the normalised random walk

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $X_1, ..., X_n$ are independent and identically distributed random variables, the tools most widely available are the weak and strong laws of large numbers, and the Central Limit Theorem. The former give us a value for which we can be reasonably certain our random walk tends to in the long run, and the latter provides us with an approximate distribution for our random walk when n is large. Both of these methods have similar limitations. The weak and strong laws of large numbers don't give any characterisation of probability relating to the behaviour of our random walk, and the Central Limit Theorem, although it attempts to, has been known to be inaccurate in calculating the probability of rare events. For example, when X_1 is Bernoulli with $p = 0.2$,

$$\mathbb{P}[S_{100} \leq 1/100] \approx 5.09 \times 10^{-9},$$

yet

$$\mathbb{P}[\mathcal{N}(1/5, (1/25)^2) \leq 1/100] \approx 3 \times 10^{-7},$$

from [1]. Although both quantities are small, the Central Limit Theorem approximation is almost 100 times the size of the actual value. (Admittedly this example is slightly cherry picked, partially to enable the use of standard Normal tables, but we will return to it later.)

1. CRAMÉR'S THEOREM IN $\mathbb{R}$

To remedy this discrepancy Harald Cramér proved what is now referred to as Cramér's Large Deviation Theorem. It has some generalisations, and has inspired similar theorems, but its original form is stated as follows.

Theorem 1. *(Cramér) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables with cumulant generating function $\Lambda : \mathbb{R} \to (-\infty, \infty]$, and let $S_n = n^{-1}\sum_{i=1}^{n} X_i$. Then*

$$-\inf_{x\in B^\circ} \Lambda^*(x) \leq \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}[S_n \in B]$$
$$\leq \limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}[S_n \in B] \leq -\inf_{x\in \overline{B}} \Lambda^*(x)$$

for any measurable set B, where B° denotes the interior of B, $\overline{B}$ its closure, and

$$\Lambda^*(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda(\theta)\}.$$

This may be a lot to take in initially, especially to readers unfamiliar with measurability, interiors and closures, or limit inferior and superiors. The important thing to take from this theorem is that for every set B for which $\mathbb{P}[S_n \in B]$ makes sense, we have an idea of the behaviour of $n^{-1}\log\mathbb{P}[S_n \in B]$ for large values of n. For example, let $\{X_n\}_{n=1}^\infty$ be Bernoulli random variables with $p = 0.2$, as before. Then

$$\Lambda^*(x) = \sup_{\theta\in\mathbb{R}}\left\{\theta x - \log(1-p+pe^\theta)\right\}.$$

It can be shown using calculus that this supremum occurs when

$$e^\theta = \frac{x(1-p)}{p(1-x)}$$

which only make sense when $x \in (0,1)$. In this case

$$\Lambda^*(x) = x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right).$$

When $x \le 0$ or $x \ge 1$ the supremum occurs by letting θ tend to minus infinity or plus infinity respectively, in which case the supremum is infinity, unless $x = 0, 1$, in which case $\Lambda^*(0) = -\log(1-p)$ and $\Lambda^*(1) = -\log p$. These last values coincide with the expression for Λ^* above, by defining $0\log 0 = 0$. So, in conclusion, we have

$$\Lambda^*(x) = \begin{cases} x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right) & x \in [0,1] \\ +\infty & \text{otherwise.} \end{cases}$$

This function is almost continuous (continuous apart from at $x = 0, 1$), so we can prove that $\inf_{x\in(a,b)}\Lambda^*(x) = \inf_{x\in[a,b]}\Lambda^*(x)$ for all $a, b \in [0,1]$ and hence that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}[S_n \in (a,b)] = -\inf_{x\in[a,b]}\Lambda^*(x).$$

Further analysis of Λ^* will reveal that it is non-negative and has a unique zero when $x = p = \mathbb{E}[X_1]$. Also, the above statement implies that, for large n,

$$\mathbb{P}\left[S_n \in (a,b)\right] \approx f(n)\exp\left\{-n\inf_{x\in[a,b]}\Lambda^*(x)\right\}$$

for some function f that increases sub-exponentially if at all, and is possibly dependant on (a,b). Under this construct, returning to our original Bernoulli

example for $p = 0.2$,

$$\mathbb{P}\left[S_n \leq \frac{1}{100}\right] \approx f(100)\exp\left(-100\left(\frac{1}{100}\log\left(\frac{1}{20}\right) + \frac{99}{100}\log\left(\frac{99}{80}\right)\right)\right)$$
$$= f(100)\left(1.38 \times 10^{-8}\right)$$

which is still significantly better than our Central Limit Theorem approximation if we assume $f \approx 1$. It can be proven easily that $f = 1$ uniformly if we examine $\mathbb{P}[S_n = 0]$, so this may not be a bad assumption to make.

Other direct observations we can make are that for two sets, say $[.1, .15]$ and $[.25, 3]$, we have

$$\lim_{n\to\infty} \frac{\log \mathbb{P}[S_n \in [.1, .15]]}{\log \mathbb{P}[S_n \in [.25, .3]]} = \lim_{n\to\infty} \frac{\frac{1}{n}\log \mathbb{P}[S_n \in [.1, .15]]}{\frac{1}{n}\log \mathbb{P}[S_n \in [.25, .3]]} = \frac{8.3786}{7.382} = 1.135.$$

So although both events become increasingly unlikely for large n, we can see that, roughly speaking, S_n is more likely to deviate far below the mean than above the mean.

Returning to the general statement of Cramér's Theorem, with some restrictions on the distribution of X_1, a number of properties of Λ^* can be proven.

Lemma 1. *Let X a random variable and let Λ^* be defined as in Cramér's Theorem. Then Λ^* is a non-negative convex lower semi-continuous extended real valued function. Moreover, if the moment generating function of X is finite in a neighbourhood of the origin, then $\overline{x} = \mathbb{E}[X]$ exists as a real number and $\Lambda^*(x) = 0$ if and only if $x = \overline{x}$, and the level sets*

$$\Psi(\alpha) = \{x : \Lambda^*(x) \leq \alpha\}$$

are compact for all $\alpha \in [0, \infty)$.

Table 7.1 lists the form of Λ^* for a number of distributions. Notice how the Lognormal and Cauchy rate functions don't have unique zeros or compact level sets, as their cumulant generating functions are not finite in a neighbourhood of the origin.

2. LARGE DEVIATION PRINCIPLES IN GENERAL

The applications of Cramér's Theorem and other theorems in Large Deviation Theory are diverse and far reaching; they can be found, for example, in statistical mechanics, thermodynamics and hypothesis testing. However they are also quite technical and often esoteric, so they will not be discussed here. Instead we will define and discuss the core definition in Large Deviation Theory, the Large Deviation Principle. First however, some preliminary definitions are required, mostly based on some material from functional analysis and general topology.

Definition 1. *If $\mathcal{X}$ is Hausdorff topological space, then a function $I : \mathcal{X} \to [0, \infty]$ is a rate function if I is lower semi-continuous, and its level sets $\Psi_I(\alpha) = \{x : I(x) \leq \alpha\}$ are compact in $\mathcal{X}$.*

Distribution	Rate Function
Bernoulli(p)	$\Lambda^*(x) = \begin{cases} x\log\left(\frac{x}{p}\right) & \\ \quad +(1-x)\log\frac{1-x}{1-p} & x \in [0,1] \\ +\infty & \text{otherwise} \end{cases}$
Poisson(λ)	$\Lambda^*(x) = \begin{cases} x\log\left(\frac{x}{\lambda}\right) - x + \lambda & x \in [0,\infty) \\ +\infty & \text{otherwise} \end{cases}$
Binomial(n, p)	$\Lambda^*(x) = \begin{cases} x\log\left(\frac{x}{p}\right) & \\ \quad +(n-x)\log\left(\frac{n-x}{1-p}\right) & \\ \quad -n\log n & x \in [0,n] \\ +\infty & \text{otherwise} \end{cases}$
Normal(μ, σ^2)	$\Lambda^*(x) = \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$
Lognormal(μ, σ)	$\Lambda^*(x) = \begin{cases} 0 & x \geq 0 \\ +\infty & \text{otherwise} \end{cases}$
Cauchy(γ)	$\Lambda^*(x) = +\infty \ \forall \ x \in \mathbb{R}$

Table 7.1: Form of the rate function Λ^* for a number of common distributions.

Note that in the case of Cramér's Theorem, Λ^* is a rate function if the underlying random variable X obeys the conditions in Lemma 1.

Definition 2. *Let $\mathcal{X}$ be some set, and let $2^{\mathcal{X}}$ be its power set. $\Sigma \subset 2^{\mathcal{X}}$ is a σ-algebra of $\mathcal{X}$ if*

1. *Σ is non-empty.*
2. *$A_1, ..., A_n, ... \in \Sigma \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \Sigma$ (Closed under countable union).*
3. *$A \in \Sigma \Rightarrow A^c \in \Sigma$ (Closed under complementation).*

Here A^c denotes the complement of A. Note that by using De Morgan's Laws, (ii) and (iii) can be used to prove that Σ is closed under countable intersection. $(\mathcal{X}, \Sigma)$ is referred to as a measurable space.

Definition 3. *Let $(\mathcal{X}, \tau)$ be some topological space. Then $\mathcal{B}(\mathcal{X}, \tau)$, the Borel σ-algebra on $(\mathcal{X}, \tau)$ is the smallest σ-algebra that contains all the open sets in $(\mathcal{X}, \tau)$. If $\mathcal{B}(\mathcal{X}) \subseteq \Sigma$ for some σ-algebra Σ, then Σ is said to contain the Borel σ-algebra on $(\mathcal{X}, \tau)$.*

σ-algebras are of huge importance in probability, especially when analysing probabilities on a sample space that isn't $\mathbb{R}^n$. When setting up a probability measure on a sample space we must do so with respect to some σ-algebra on the space, often the Borel σ-algebra. With these definitions in mind we are now ready to define the Large Deviation Principle.

Definition 4. *Let $(\mathcal{X}, \tau)$ be a Hausdorff topological space, let Σ be a σ-algebra that contains $\mathcal{B}(\mathcal{X})$, let $\{\mathbb{P}_n[\cdot]\}_{n=1}^{\infty}$ be a sequence of probability measures on $(\mathcal{X}, \Sigma)$, and let $\{a_n\}_{n=1}^{\infty}$ be a positive, strictly increasing sequence tending to infinity. Then the pair $\{(\mathbb{P}_n[\cdot], a_n)\}_{n=1}^{\infty}$ is said to obey a Large Deviation Principle with rate function $I : \mathcal{X} \to [0, \infty]$ if*

$$-\inf_{x \in B^\circ} I(x) \leq \liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n[B] \leq \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n[B] \leq -\inf_{x \in \overline{B}} I(x)$$

for every $B \in \Sigma$. If $a_n = n$ for all $n \geq 1$ then we say $\{\mathbb{P}_n[\cdot]\}_{n=1}^{\infty}$ obeys a Large Deviation Principle with rate function I.

Note that by this definition, Cramér's Theorem states that $\{\mathbb{P}[S_n \in \cdot\,]\}_{n=1}^{\infty}$ obeys a Large Deviation Principle with rate function Λ^* if the underlying random variables $\{X_n\}_{n=1}^{\infty}$ have a moment generating function that is finite in a neighbourhood of the origin. Also, just like in Cramér's Theorem, unique zeros of the rate function are sufficient to prove weak laws, regardless of the properties of the topological space.

3. EXAMPLES OF LARGE DEVIATION PRINCIPLES

There are a number of different methods of proving Large Deviation Principles, from direct proofs, many examples of which can be found in a text by Amir and Dembo on Large Deviation Theory [2], to general methods of proof compiled by Lewis and Pfister [3]. However, we will summarise by looking at examples of Large Deviation Principles.

3.1. *Cramérs Theorem in* $\mathbb{R}^d$

Cramér's Theorem in $\mathbb{R}^d$ is simply a generalisation of Cramér's Theorem to independent and identically distributed random vectors.

Theorem 2. *(Cramér's Theorem in $\mathbb{R}^d$) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ with cumulant generating function $\Lambda : \mathbb{R}^d \to (0, \infty]$ which is finite in a neighbourhood of the origin. Then $\{\mathbb{P}[S_n \in \cdot]\}_{n=1}^{\infty}$ obeys a Large Deviation Principle with rate function*

$$\Lambda^*(\boldsymbol{x}) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, \boldsymbol{x} \rangle - \Lambda(\lambda)\},$$

where $\langle \lambda, \boldsymbol{x} \rangle$ is the inner product between the vectors λ and $\boldsymbol{x}$.

3.2. *Gärtner-Ellis Theorem*

The Gärtner-Ellis Theorem is a generalisation of Cramér's Theorem, where the assumption that $\{X_n\}_{n=1}^{\infty}$ are i.i.d. is dropped. We will state the theorem in the case where $X_n \in \mathbb{R}^d$ for all n. Let Λ_n be the cumulant generating function of X_n. The i.i.d. nature of the sequence of random vectors is replaced with the following assumption.

Assumption 1. *For each* $\lambda \in \mathbb{R}^d$,

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda)$$

exists as an extended real number. Furthermore, Λ *is finite in a neighbourhood of the origin.*

Notice how this assumption holds trivially when $\{X_n\}_{n=1}^{\infty}$ are i.i.d.

Definition 5. $y \in \mathbb{R}^d$ *is an exposed point of* $\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, \boldsymbol{x}\rangle - \Lambda(\lambda)\}$ *if for some* $\lambda \in \mathbb{R}^d$ *and all* $x \neq y$,

$$\langle\lambda, \boldsymbol{x}\rangle - \Lambda^*(y) > \langle\lambda, \boldsymbol{x}\rangle - \Lambda^*(x).$$

Definition 6. *Let* $\mathcal{D}_\Lambda = \{\lambda \,|\Lambda(\lambda) < \infty\}$ *and let* $\mathcal{D}_\Lambda^\circ$ *be its interior.* Λ *is essentially smooth if*

1. $\mathcal{D}_\Lambda^\circ \neq \emptyset$,
2. Λ *is differentiable on* $\mathcal{D}_\Lambda^\circ$,
3. $\lim_{n\to\infty} |\nabla\Lambda(\lambda_n)| = \infty$ *for all sequences* $\{\lambda_n\}_{n=1}^{\infty}$ *in* $\mathcal{D}_\Lambda^\circ$ *converging to a boundary point of* $\mathcal{D}_\Lambda^\circ$.

Then the Gärtner-Ellis Theorem is stated as follows.

Theorem 3. *(Gärtner-Ellis) Let Assumption 1 hold with a function* Λ *that is essentially smooth and lower semi-continuous. Then* $\{\mathbb{P}[S_n \in \cdot]\}_{n=1}^{\infty}$ *obeys a Large Deviation Principle in* $\mathbb{R}^d$ *with rate function*

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, \boldsymbol{x}\rangle - \Lambda(\lambda)\}.$$

3.3. Sanov's Theorem

Theorem 4. *(Sanov's Theorem) Let* $\mathcal{X}$ *be a complete seperable metric space, let* $\{X_i\}_{i=1}^{\infty}$ *be a sequence of random variables in* $\mathcal{X}$ *with distribution* μ. *Let* $\mathcal{M}(\mathcal{X})$ *be the space of probability measures on* $\mathcal{X}$, *and let* τ *be the topology on* $\mathcal{M}(\mathcal{X})$ *generated by a base consisting of all sets of the form*

$$U_{x,\delta,f} = \left\{\nu \middle| \left|\int_{\mathcal{X}} f d\nu - x\right| < \delta\right\}$$

for $x \in \mathbb{R}$, $\delta > 0$, $f \in \mathcal{X} \to \mathbb{R}$ *bounded and continuous. Let*

$$L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

be the empirical measure corresponding to the first n observations, δ_{X_k} *denoting the unit mass at* X_k. *Then* $\{\mathbb{P}[L_n \in \cdot]\}_{n=1}^{\infty}$ *obeys a Large Deviation Principle with rate function*

$$\mathcal{H}(\nu||\mu) = \int_{\mathcal{X}} d\nu \log\frac{d\nu}{d\mu}$$

So in the same way that Cramer's Theorem served to estimate probabilities of sample averages, Sanov's Theorem estimates probabilities of sample densities. The topology referred to above is known as the weak topology. For example, if $\{X_n\}_{n=1}^{\infty}$ are from a Normal(γ, σ^2) distribution, and ν represents an Exp(λ) distribution. Then

$$d\mu = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\gamma)^2}{\sigma^2}} dx \text{ and } d\nu = \lambda e^{-\lambda x} dx$$

and so

$$\begin{aligned}
\mathcal{H}(\nu||\mu) &= \int_0^{\infty} \lambda e^{-\lambda x} \log \left(\frac{\lambda e^{-\lambda x}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\gamma)^2}{\sigma^2}}} \right) dx \\
&= \int_0^{\infty} \lambda e^{-\lambda x} (\log(\lambda\sigma\sqrt{2\pi}) + \frac{(x-\gamma)^2}{\sigma^2} - \lambda x) dx \\
&= \log(\lambda\sigma\sqrt{2\pi}) + \frac{1}{\sigma^2} \left(\gamma^2 + \frac{2}{\lambda^2} - \frac{2\gamma}{\lambda} - \sigma^2 \right).
\end{aligned}$$

So the probability of obtaining what "looks like" an exponential distribution from a sample of normal random variables decays exponentially with the above parameter. Using calculus we can show that when $\gamma = 0, \lambda = \frac{2}{\sigma}$ minimises the above expression, and therefore out of all exponential distributions, an empirical sample from a Normal$(0, \sigma^2)$ is most likely to converge to one with mean $\frac{\sigma}{2}$.

BIBLIOGRAPHY

[1] http://web.pdx.edu/ stipakb/download/PA551/NormalTable.gif

[2] Large Deviation Techniques and Applications, Amir Dembo and Ofer Zeitouni

[3] Thermodynamic Probability Theory: Some Aspects of Large Deviations.

SIMULATING A TSUNAMI IN MATLAB

ANTHONY JAMES MCELWEE
University College Dublin

1. INTRODUCTION

This article presents a simulation that depicts a tsunami wave crossing an ocean. It adapts a finite-difference finite-time scheme that was implemented using the Lax-Wendroff method in order to solve the Shallow Water Wave Equations. While computational methods are time consuming and often difficult to implement, the rewards for implementing such schemes are high and often visually pleasing. Indeed, there is a general feeling nowadays that properly trained modern applied mathematicians, physical scientists, and engineers should have some understanding of numerical methods. The computational work was coded and rendered in Matlab. This software is available to most undergraduate students in science and engineering disciplines. Upon reading this article, it is hoped the reader will be able to create an .avi film clip that depicts the a basic simulation of a tsunami that is at the far-from-shore stage of propagation. Screen grabs of such a film are pictured below.

2. PREVIOUS WORK

This simulation builds on previous work carried out by Moler [3]. While this code creates a simulation pictured below, it lacks graphical quality and is incapable of being directly captured in a film file.

Moler's use of special loops for stability and graphics initialisation are common coding practice but such elements make the code inaccessible to undergraduates who want to play with the finite-difference finite-time schemes and the visual parameters of the simulation. Any change could prevent the simulation from running and thus serve as a disincentive to students investigating numerical schemes. The new simulation created a dramatically altered code that is geared towards including easily adaptable graphical parameters and the ability to be captured to an .avi file. In order to emphasis these aspects, the new code abandoned much of the numerical scheme. However, a partial derivation of the numerical scheme from the partial differential equations will be outlined in the next section.

Mathematical Scheme

The Shallow Water Waves imitate waves whose depth is much less than the wavelength of the disturbance of the free surface. The model assumes that the vertical acceleration of the fluid is much smaller than the gravitational acceleration and that the horizontal component of the fluid velocity is uniform along any vertical section through the fluid (Billingham and King [1]). The

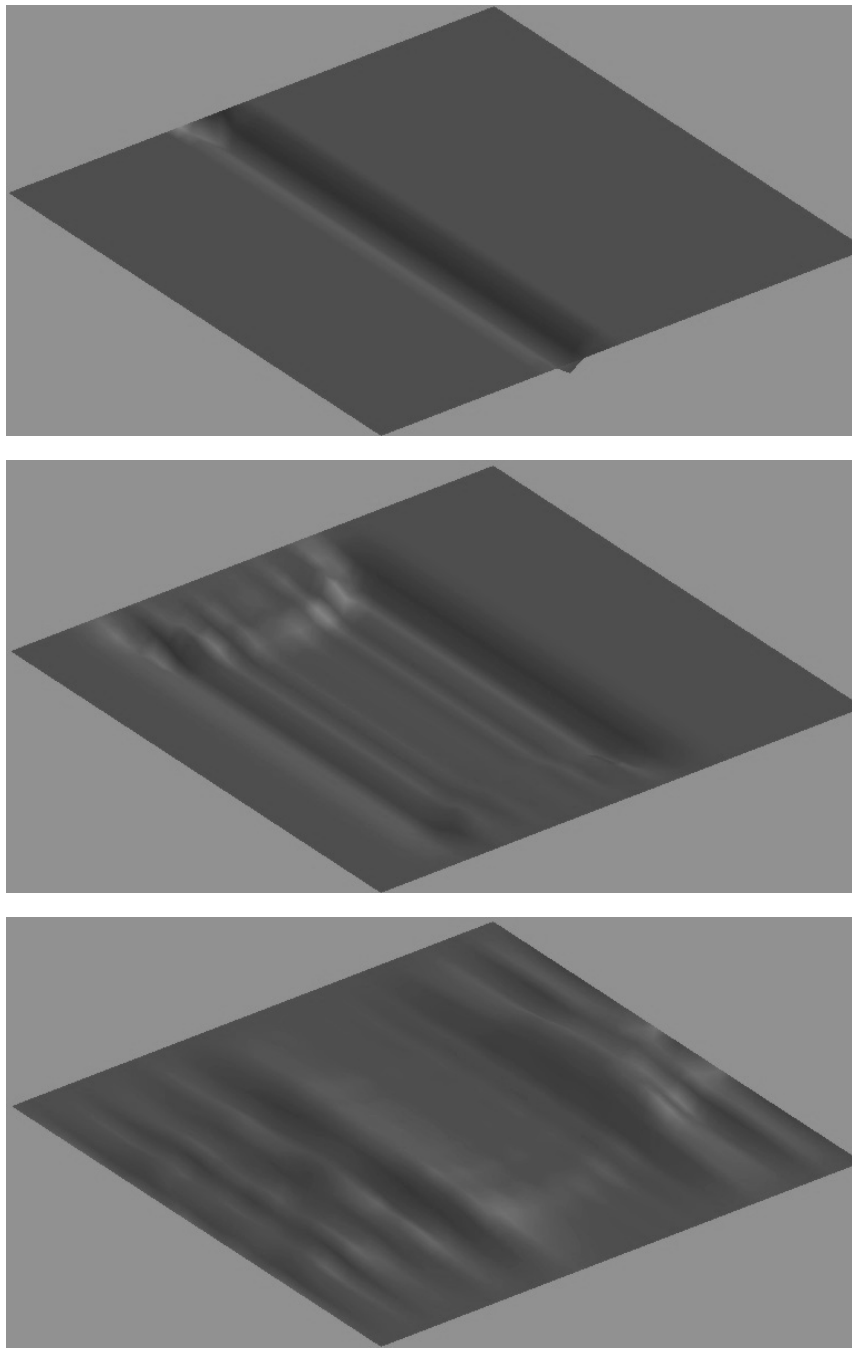


Figure 8.1: Screen grabs of the final simulation.

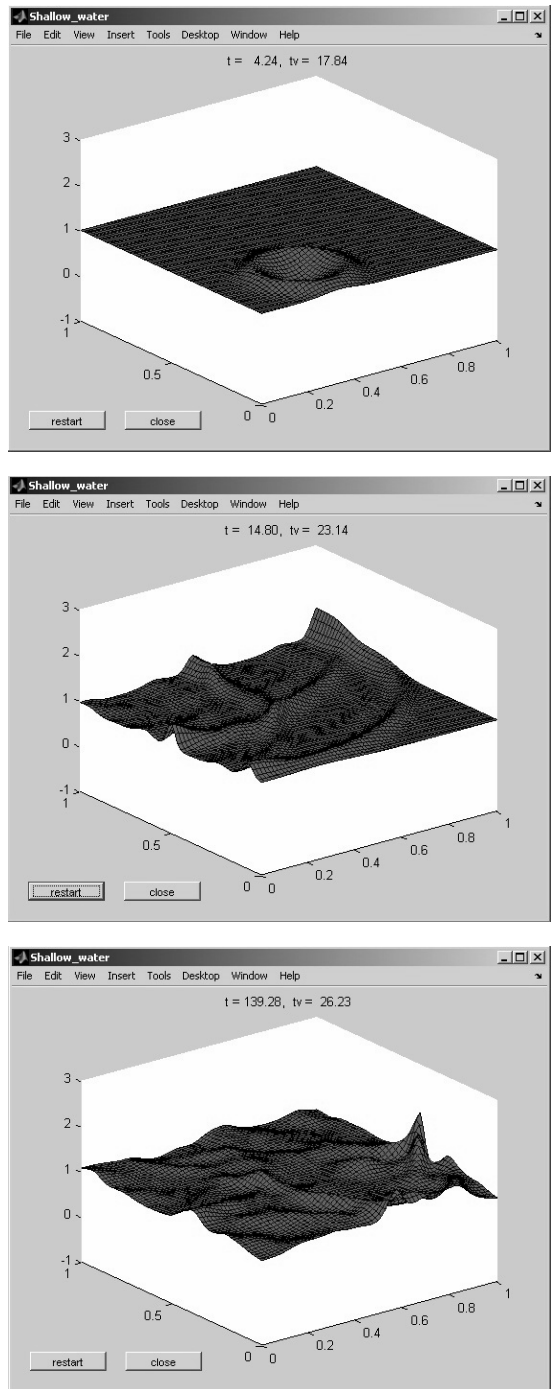


Figure 8.2: Screen grabs of Moler's simulation.

partial differential equations used by Moler [3] correspond to those derived by Billingham and King [1] from the Navier-Stokes equations. These derived equations are

$$\begin{aligned}
\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + \frac{\partial (vh)}{\partial y} &= 0 \\
\frac{\partial (uh)}{\partial t} + \frac{\partial (u^2 h + \frac{1}{2} g h^2)}{\partial x} + \frac{\partial (uvh)}{\partial y} &= 0 \\
\frac{\partial (vh)}{\partial t} + \frac{\partial (uvh)}{\partial x} + \frac{\partial (v^2 h + \frac{1}{2} g h^2)}{\partial x} &= 0
\end{aligned}$$

where h corresponds to the height of the water and g is the gravitational acceleration. The two-dimensional velocity field is denoted in terms of u and v and their products with h are proportional to the water's momentum. The perspective taken is that the mass and momentum of the water is conserved at all times. A number of numerical difference methods could be considered that are suitable for conservation problems but the Lax-Wendroff method is satisfactory for this situation.

The next step is to recast the equations so that they are in the form of a hyperbolic partial differential equation which obeys the conservation laws. To do this, the following three vectors were formed:

$$U = \begin{pmatrix} h \\ uh \\ vh \end{pmatrix}, \qquad F(U) = \begin{pmatrix} uh \\ u^2 h + \frac{1}{2} g h^2 \\ uvh \end{pmatrix}, \qquad G(U) = \begin{pmatrix} vh \\ uvh \\ v^2 h + \frac{1}{2} g h^2 \end{pmatrix},$$

which gives our model equation

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} + \frac{\partial G(U)}{\partial y} = 0.$$

The U corresponds to the height matrix, while F is the matrix that stores the zonal velocity and G stores the meridional velocity.

Numerical Scheme

Unfortunately, any land masses or irregular geometry have to be discounted in order to keep the scheme relativity simple. This implementation is over a square grid with regular divisions. Such a grid is the easiest to implement in Matlab. The grid is also stepped so that time and space are subdivided, as pictured in Moler's brief guide to the original code.

Such a grid can be revealed by altering the code provided later in article so that `Edge_Alpha` has a value close to 1, rather than 0.

As already stated, the original code uses the Lax-Wendroff numerical scheme to solve the new model hyperbolic partial differential equation. The

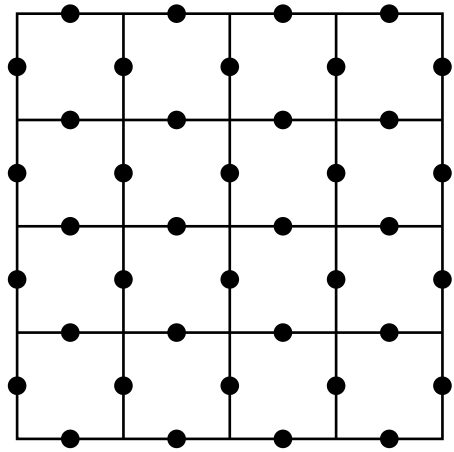


Figure 8.3: The grid showing the values for the vectors positioned in a half step position on both axis.

steps involved in the scheme are described in most computational fluid dynamics books and freely available on the internet. To quickly summarise, the finite difference representation of the model equation is derived from the two dimensional Taylor series expansion of the dependant variables. The time derivative of the model equation is taken to obtain a new equation and then the model equation and the new equation are substituted back into the Taylor expansion of the dependant variables. In one direction, say x to illustrate, the equation becomes

$$U_i^{n+1} = U_i^n - \triangle t \left[\frac{U_{i+1}^n - U_{i-1}^n}{2\triangle x} \right] + \frac{1}{2}(\triangle t)^2 \left[\frac{U_{i+1}^n - 2U_i^n + U_{i-1}^n}{(\triangle x)^2} \right].$$

Moler goes further and half steps the time and space steps.

The full set of equations for the two time steps are available in Moler [3]; the present author sees no benefit in rehashing them here. At this point we depart from Moler and decide to take action with a view to improving the visual quality of the simulation. For the simulation will assume that only the y-components of the equations matter and set the x-components equal to the y-components. This shortens the code and since this isn't a real tsunami no one should really care! The code could be rectified by the reader to fully reflect the nature of the Shallow Water Waves. Provided the stability of the numerical scheme is maintained and the film looks like a wave propagating across a large body of water, the goal of this exercise will have been achieved. This allows us to get into the visual aspect of the simulation.

3. METHODOLOGY

The adapted code is now explained in the order that it appears in the programme. The reader should try adapting the code parameters further and experiment with the visual effects. If any of the code remains unclear after reading the explanations, extra help can be found using the help function in Matlab.

The first thing to do is to declare the function in a new .m file that is saved in the Matlab workspace. This allows the code to be easily updated. Here the function is called `FDFTSWWIUMM`, (Finite Difference Finite Time Shallow Water Waves Irish Undergraduate Mathematical Magazine). Clearing commands are included so that when the simulation is restarted, the old values will not interfere with the fresh simulation.

```
1  function FDFTSWWIUMM
2
3  %% Clear MATLAB
4  clear all;clc;clf
```

The next section establishes some basic scene settings. The axes are stripped away and the simulation freezes its aspect ratio, otherwise the simulation will lack any sort of aesthetic appeal. The z-axis, corresponding to the height of the water, is also fixed. A failure to set the z-axis would result in the scene having a jittery appearance, even if the amplitude of the perturbation wave is relatively small.

Setting the handle for the background will probably be new to the reader. This gcf is setting the background colour using the RGB colour code with a base of 255 integer values. The surface handle is set in a similar manner. The `water_range` generates a vector for the surface colour.

```
5  %% Scene Settings
6  axis vis3d off
7  zlim([-50 50])
8  set(gcf,'color',[57/255 176/255 121/255]);
9  water_range = [20/255 80/255 118/255];
```

The next step is to set any constants such as gravity, the time step and the number of time steps the simulation will last. Initially, the time length is set at 40. It is kept short since a requirement to change some basic parameter can be spotted by the reader in that time frame, removing the need to repeatedly run ling simulations.

```
10 %% Constants
11 grav = 9.81;
12 dt = 0.5;
13 sim_steps=40;
```

The next section declares a batch of variables that are the handle property settings for surfaces. This is not a complete list of possible variables, but they are comprehensive for this type of simulation. They will define the handle's set properties. In order to make it easier to adapt the variables, they have been named in a similiar fashion to those property titles. This allows you to

leave the surface handle commands alone and therefore reduce the chances of accidentally introducing an error into the code.

```
%% Set Handle Property Settings.
Ambient_Strength=0.5;
BackFace_Lighting='lit';
Diffuse_Strength=0.8;
Edge_Alpha=.1;
Edge_Color=water_range;
Edge_Lighting='phong';
Erase_Mode='normal';
Face_Alpha=0.9;
Face_Color=water_range;
Face_Lighting='phong';
Mesh_Style='both';
Normal_Mode='auto';
SpecularColor_Reflectance=.5;
Specular_Exponent=10;
Specular_Strength=.5;
```

The section above offers a lot of possibilities and is the key to enhancing the visual properties of the simulation. The obvious place to start is by varying the numbers but some of the settings are set by string commands. More information about the other available options for these string commands are available at www.mathworks.co.uk/help/matlab/ref/surface_props.html. For example, `gouraud` shading could be used instead of the `phong` settings. These commands have a vast amount of physical attributes that go to the core of computer graphics. By varying these properties, the simulation can take on a new appearance.

The next section is a straightforward creation of the square, uniformly divided grid. Instability will occur if the grid step and the time step are beyond certain values. This is manifested in a simulation that fails to run, or blows up rapidly. Usually, the parameters can be choosen before running the simulation. For example, the Lax-Wendroff scheme usually has the stability condition that corresponds to the time step over the spatial step having a ratio less than one. Since a lot of the numerical scheme has been dumped, no definition of the stability for the new situation exists. The reader should experiment by changing these variables to understand the significance of selecting step sizes.

```
%% Grid Parameters
X_length = 100;
Y_length = X_length;
X_n = 35;
Y_n = X_n;
dx = X_length/(X_n-1);
dy = dx;

%% Grid Generation
[x_grid] = 0:dx:X_length;
[y_grid] = 0:dy:Y_length;
```

The next section sets up the initial conditions of the surface so that the ocean is calm and flat. One could add some noise to the surface to mimic the normal action of the sea but due to the scale that tsunamis are modelled

on, such noise based disturbance would be pointless. There were boundary conditions included in the original code that were perfectly reflective and so that the waves bounced off the sides of the grid. This is not the behaviour of a real tsunami and since some of the numerical scheme was thrown away the author has decided to get rid of the boundary conditions too. The reflective nature of the boundaries in the new code is attributed to the butchered scheme outlined later in the article.

```
41 %% Initial Conditions
42 height = zeros(X_n,Y_n);height_x = zeros(X_n,Y_n);height_y
       = zeros(X_n,Y_n);
43 zonal = zeros(X_n,Y_n);zonal_x = zeros(X_n,Y_n);zonal_y =
       zeros(X_n,Y_n);
44 meridional = zeros(X_n,Y_n);meridional_x = zeros(X_n,Y_n);
       meridional_y = zeros(X_n,Y_n);
```

Now the simulation needs a perturbation on the surface so that we can get some action into the simulation. A tsunami-like disturbance can be created which propagates in time and space by plotting a simple ridge uniformly across the width of the surface. The author recommends that the reader tries different plots and perturbations. For example, the use of a circle could mimic the impact of an asteroid hitting the ocean.

```
45 %% Pertubation Of Surface [Tsunami Wave]
46 for i=1:X_n
47     Ts = 1;
48     wavelength = 1000;
49     amplitude=5;
50     phase=((((X_n-1)/(wavelength))-(dt/Ts))+(((Y_n-1)/(
           wavelength))-(dt/Ts)));
51     height(i,15)=amplitude*sin(2*pi*(1/Ts)+phase);
52 end
```

Before the time progression of the simulation is implemented, Matlab needs to be able to capture and export the simulation in an .avi file. The full path is set to where the simulation will be exported. The `gca` command shown is included so that the axis are returned to each frame captured by Matlab. If the `gca` command is omitted, the axis settings will default every time the loop executes and Matlab will stall almost immediately.

```
53 %% Movie Capture Part01
54 vidObj = VideoWriter('C:\Desktop\latest1.avi');
55 open(vidObj);
56
57 set(gca,'nextplot','replacechildren');
```

The numerical part of the programme, which will dictates how the surface behaves on the screen, will now be dealt with briefly. The scheme is looped for the designated number of time steps. For each time step, all of the matrices are recalculated using the previous values and the following scheme:

$$
\begin{aligned}
Hy_{(i,j)} &= 2H_{(i,j)} - Hx_{(i,j)} + \left(\frac{2dt^2}{dxdy}\right)\{H_{(i+1,j)} \\
&\quad + H_{(i-1,j)} + H_{(i,j+1)} + H_{(i,j-1)} - 4H_{(i,j)}\}, \\
Fy_{(i,j)} &= Fx_{(i,j)} - g\frac{dt}{dx}(H_{(i+1,j)} - H_{(i-1,j)}), \\
Gy_{(i,j)} &= Gx_{(i,j)} - g\frac{dt}{dy}(H_{(i,j+1)} - H_{(i,j-1)}).
\end{aligned}
$$

The loop sets the next step values of the x-axis values to the current height values and then sets the next step values of the height values to the current y-axis values. The result holds no real mathematical significance. It is merely a happy accident but the scheme works for our purposes.

```
%% Time Progression
for n=1:sim_steps
    for i=2:X_n-1
        for j=2:Y_n-1
    zonal_y(i,j)=zonal_x(i,j)-grav*(dt/dx)*(height(i+1,j)
        -height(i-1,j));
    meridional_y(i,j)=meridional_x(i,j)-grav*(dt/dy)*(
        height(i,j+1)-height(i,j-1));
    height_y(i,j)=(2*height(i,j)-height_x(i,j)+(((2*dt^2/
        dx*dy))*(height(i+1,j)+height(i-1,j)+height(i,j
        +1)+height(i,j-1)-4*height(i,j))));
        end
    end
 height_x=height;
 height=height_y;
```

The next section gives the code for the surface handle and its properties. Unless the reader wants to insert a new property, it is best to leave this section alone and just change the variables previously declared.

```
surf_handel = surf(x_grid,y_grid,height);
    set(surf_handel,...
        'AmbientStrength',Ambient_Strength,...
        'BackFaceLighting',BackFace_Lighting,...
        'DiffuseStrength',Diffuse_Strength,...
        'EdgeAlpha',Edge_Alpha,...
        'EdgeColor',Edge_Color,...
        'EdgeLighting',Edge_Lighting,...
        'EraseMode',Erase_Mode,...
        'FaceAlpha',Face_Alpha,...
        'FaceColor',Face_Color,...
        'FaceLighting',Face_Lighting,...
        'MeshStyle',Mesh_Style,...
        'NormalMode',Normal_Mode,...
        'SpecularColorReflectance',
            SpecularColor_Reflectance,...
        'SpecularExponent',Specular_Exponent,...
        'SpecularStrength',Specular_Strength)
```

This final part defines and positions the global lighting source for the scene. The generated frame gets pushed to the video object where it is stored and

finally closed. Now run the whole programme from your script environment and it's lights, camera, action!

```
light('Position',[100 40 80],'Style','infinite');

 %% Camera Position
 campos('auto')

%% Movie Capture Part02
    currFrame = getframe;
    writeVideo(vidObj,currFrame);
 end

%% Movie Capture Part03
close(vidObj);
```

4. CONCLUSIONS

A simulation has been produced that looks like a tsunami wave propagating across an ocean, even if the readers know that it is not a proper representation of the Shallow Water Waves. By varying the properties, the reader can alter the visual output of the scene and change how the scene is rendered by Matlab. Some simple examples are shown below.

The author recommends Moler [3] for some further information on the original code that provided the ground work of this article. A future project could be to tackle a full implimentation of the Shallow Water Waves using Lax-Wendroff, or using the MacCormack method, in combination with the surface handle and .avi export options. The use of a moving camera should also be a investigated. Those interested in more advanced material in the area of modelling large sea surfaces should consult the freely available Løset [2].

BIBLIOGRAPHY

[1] J. Billingham and A.C. King. *Wave Motion*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2000. ISBN 9780521634502. URL `http://books.google.ie/books?id=bNePaHM20LQC`.

[2] Tarjei Kvamme Løset. *Real-Time Simulation and Visualization of Large Sea Surfaces*. PhD thesis, Norwegian University of Science and Technology, 2007.

[3] C.B. Moler. waterwave file, January 2012. URL `www.mathworks.co.uk/moler/exm/exm/waterwave.m`. Accessed 14/10/2012.

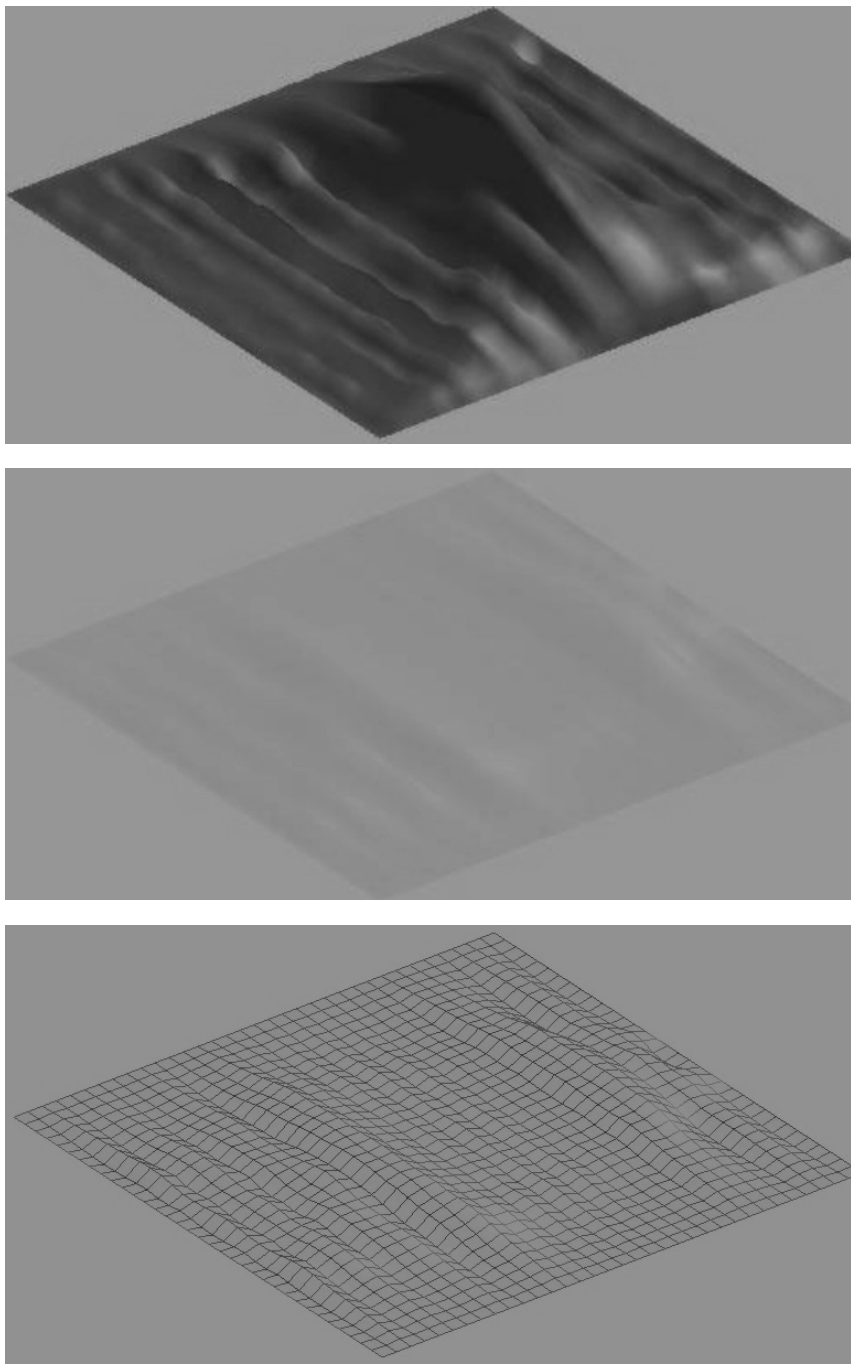


Figure 8.4: (a) Increased initial amplitude. (b) Reducing `Face_Alpha` and increasing the Ambient Strength. (c) Removing the cell faces and plotting the mesh.

COMPUTATIONAL ANALYSIS OF THE DYNAMICS OF THE DIPPY BIRD

GLENN MOYNIHAN AND SEÁN MURRAY

Trinity College Dublin

A simple, quantitative, model was constructed and solved in *Mathematica* that replicated the short-, medium- and long-term motion of the Dippy Bird (DB). An extended model for two coupled DBs was then constructed and produced motion that displayed stable in-phase and anti-phase motion as well as period doubling for the second bird. Bifurcation diagrams were produced in an attempt to observe chaotic behaviour for which we found some signs.

1. INTRODUCTION

The DB is an entertaining toy that has adorned the desks of enthusiasts since its invention in the early 20^{th} century. Much of its appeal comes from its apparent display of perpetual motion; however, upon a closer look we can see that the bird is in fact a thermal engine that is obeying simple thermodynamical principles.

The source of the DB's power comes from the temperature difference induced by the evaporation of water on the bird's head. Figure 9.1 is a schematic diagram of the bird outlining the basic mechanism. The bird consists of two glass bulbs (head and body), connected by a thin glass tube (neck) that is submerged in a volatile liquid in the body. The liquid and its vapor are all that reside within the bird - there is no air present. The glass ensemble is mounted on a pivot that allows it to rotate backwards and forwards and the entire system is in thermal equilibrium with the environment.

The bird begins tilted at an angle (near the upright position) and when released starts to oscillate. The head is covered in felt that retains water. If the humidity of the environment is less than 100%, spontaneous evaporation of the water occurs. As the water evaporates it absorbs the latent heat of vaporisation from the head, thereby lowering the temperature of the head. This decrease in temperature causes the pressure of the vapor in the head to drop according the Clausius-Clapyron relation. The pressure drop forces the liquid to rise up the neck thus raising the bird's centre of mass. As the centre of mass transitions to the head the amplitude of the oscillation decreases and the bird simultaneously begins to fall further forward with each swing. At a critical angle (near the horizontal position) an air channel forms in the neck allowing the pressure between the head and body to equalise (image on the right in figure 9.1). The liquid rushes back to the body, lowering the centre of mass and beginning the next cycle.

The DB possesses three intrinsic time-scales: the first (short) time-scale is the time taken for each single oscillation, the second (medium) timescale is that between successive "dips" where the cycle resets and the third (long) timescale is one that demonstrates the increase in the cycle period. The rate of evaporation of the water is proportional to the mass of water present on the

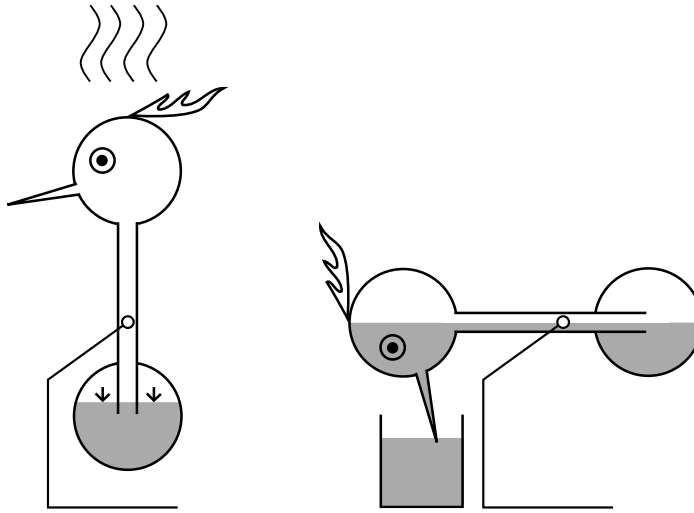


Figure 9.1: Schematic diagram of the DB in the upright and horizontal positions. In the right-hand image we note the protrusion of the neck from the liquid allowing for pressure equalisation.

head. As the water diminishes its evaporation rate slows so the whole cycle takes longer to complete. We aim to reproduce each of these time-scales in our model.

2. MOTIVATION

The motivation behind this project was to arrive at a simple, qualitative model of the DB that would describe the general physics of motion without the need for an exhaustive approach. Our initial results will be compared to that of two papers (Lorenz [1] and Guemez [2]) that investigated the DB more extensively in order to highlight the successes and shortcomings of our own model.

Lorenz [2] constructed a detailed computational model based on the thermodynamics of the bird and acquired the relevant physical parameters from experiment. Guemez [1] employed an experimental approach and using electronic devices measured the dependence of the period of the DB on environmental conditions.

3. OUR MODEL

Single Dippy Bird

The DB is, in essence, a damped, compound pendulum whose mass varies with time due to the migration of an internal liquid. However, in order to preserve the simplicity of the project we omitted the complicated evolution of the internal liquid in favour of a time-dependent, external driving-force that

would substitute nicely for the influence of the changing mass, torque and moment of inertia that would have made our model grossly more complicated and similar to that of Guemez [2].

We constructed a model based on a damped, driven pendulum and, by equating the torques arising from rate of change of angular momentum, the mass of the bird, the damping force and the driving force, we arrived at our equation of motion

$$\frac{d^2}{dt^2}\left(ml^2\theta(t)\right) + mgl\sin(\theta(t)) + bl^2\frac{d}{dt}\theta(t) - lF(t,\theta) = 0 \tag{9.16}$$

where m is the mass of the pendulum, l the (constant) string length, b the damping coefficient, θ the angle between the string and the vertical and F is the external driving force.

We now move to non-dimensionalise the equation by introducing a dimensionless time variable

$$\tau = t\sqrt{\frac{g}{l}} = \omega_0 t.$$

Then

$$\frac{d}{dt} = \frac{d}{d\tau}\frac{d\tau}{dt} = \sqrt{\frac{g}{l}}\frac{d}{d\tau}, \qquad \frac{d^2}{dt^2} = \frac{g}{l}\frac{d^2}{d\tau^2}.$$

Subbing these into equation (9.16) we arrive at

$$\frac{d^2\theta}{d\tau^2} + \sin(\theta) + \frac{b}{m\omega_0}\frac{d\theta}{d\tau} - \frac{F(\tau,\theta)}{mg} = 0. \tag{9.17}$$

We now define two control variables

$$\alpha = \frac{F(\tau,\theta)}{mg}, \qquad \beta = \frac{b}{m\omega_0},$$

yielding

$$\ddot{\theta} + \sin(\theta) + \beta\dot{\theta} - f(\tau,\theta) = 0. \tag{9.18}$$

where the driving force, f, was tailored such that it would replicate the influence of the elements we chose to neglect. It was determined *a posteriori* and set to be

$$f(\tau,\theta) = \begin{cases} \alpha\cdot(\tau-\tau_0), & \theta < \theta_{max} \\ 0, & \theta \geq \theta_{max}. \end{cases} \tag{9.19}$$

It takes the form of a sawtooth function that increases linearly until the maximum angle is reached where it resets to zero.

Including an increasing force ensures that, initially, the pendulum is allowed oscillate as normal but as time progresses the mass is pushed towards higher angles. Once it reaches the maximum angle the force suddenly cuts off (simulating the equalisation of pressure in the DB and the lowering of the centre of mass) thus restarting the cycle.

As time progresses the rate of increase of force, α, should diminish. This is to simulate mass evaporation in the physical DB and, in order to comply with this effect, we set α to decay exponentially with time; thus

$$\alpha \sim e^{-p\tau},$$

where $p \ll 1$ is our third input parameter.

We tested our model for various input parameters. Equation (9.18) was solved using the `NDSolve` function in *Mathematica* and values of θ, ω and f were recorded at each step in order to produce plots of the motion.

Coupled Dippy Birds

We extended our previous model to that of two coupled, damped, driven harmonic oscillators described in figure 9.2.

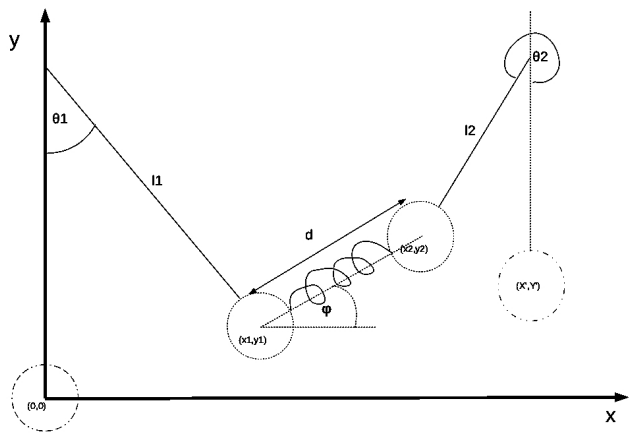


Figure 9.2: Sketch diagram of coupled DB system with the relevant physical parameters labeled. The origin is taken to be the position of Bird 1 mass in equilibrium.

We derived the following equations of motion:

$$\frac{d^2}{dt^2}(l_1^2\theta_1) + m_1 g l_1 \sin(\theta_1) + b_1\dot{\theta}_1 - l_1 F_1(\tau,\theta_1) - c l_1 \Delta d \cos(\theta_1 - \phi) = 0,$$

$$\frac{d^2}{dt^2}(l_2^2\theta_2) + m_2 g l_2 \sin(\theta_2) + b_2\dot{\theta}_2 - l_2 F_2(\tau,\theta_2) + c l_2 \Delta d \cos(\phi - \theta_2) = 0,$$

where all the symbols have their usual meanings: c is the coupling constant, Δd is the displacement of the spring from equilibrium and φ is the angle of

the spring with respect to the horizontal. The last term in each equation is the torque produced by the spring. The rest length of the spring is defined as the distance between the two masses as they hang in stable equilibrium, $d_0 = \sqrt{X^2 + Y^2}$.

Non-dimensionalising our new equations in a similar manner to before we arrive at

$$\ddot{\theta}_1 + \sin(\theta_1) + \beta_1\dot{\theta}_1 - \alpha_1 - \gamma_1 \Delta d \cos(\theta_1 - \phi) = 0, \tag{9.20}$$

$$\ddot{\theta}_2 + \sin(\theta_2) + \beta_2\dot{\theta}_2 - \alpha_2 - \gamma_2 \Delta d \cos(\phi - \theta_2) = 0, \tag{9.21}$$

where

$$\gamma_i = \frac{c}{g m_i} \text{ for } i = 1, 2.$$

Equations (9.20) and (9.21) were simultaneously solved using `NDSolve` and calculations of the same parameters were made as in the single DB case.

4. RESULTS: SINGLE DIPPY BIRD CASE

We now demonstrate the results of our model by comparing it to those obtained by Lorenz [1] and Guemez[2].

Angular Evolution

Our first goal is to highlight the qualitative aspects of our plots in figures 9.3a and 9.3b that compare to the computational-based figure 9.4, taken from [2], and the experimental-based figure 9.5 taken from [1].

- We first notice the effect of damping leading to a decrease in amplitude with each successive oscillation of the DB. This comes as no surprise as, for the first few oscillations, the force is quite small. This effect is also seen in figure 9.4.
- We next observe the shift in the equilibrium position as the force becomes dominant. This drives the angle of the bird to higher angles (or in the case of Lorenz, lower angles).
- The third, less apparent, trait is the lengthening of the period of oscillations; i.e., the third time scale. To highlight this effect we set $p = 0.002$ and plotted three successive cycles in Figure 9.3b. When compared to Figure 9.5b we see the general structure of the plots is quite similar, however our period lengthening is much more obvious.

Increasing Period

We then plotted the period of oscillations versus the cycle number for 14 successive cycles. The results are plotted in figure 9.6. We see that the slope is quite linear in contrast to figure 9.7 of Guemez [2] where we see relatively flat curve and then a sharp rise. We feel the general behaviour is comparable,

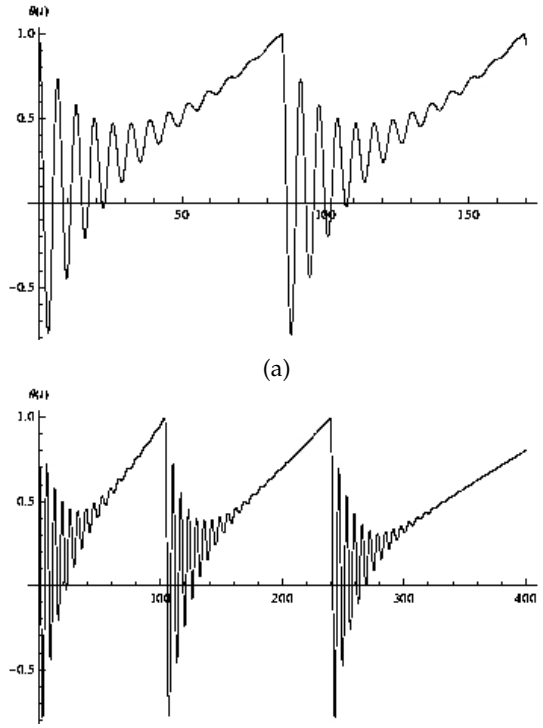


Figure 9.3: Angular evolution of DB. In both plots the x-axis marks time and the y-axis angular displacement of the bird from the vertical position of rest. Figure (a) has $p = 0$ and shows two stable cycles while figure (b) has $p = 0.002$ and shows three cycles with increasing period. We thus note the high sensitivity of the model on the parameter p.

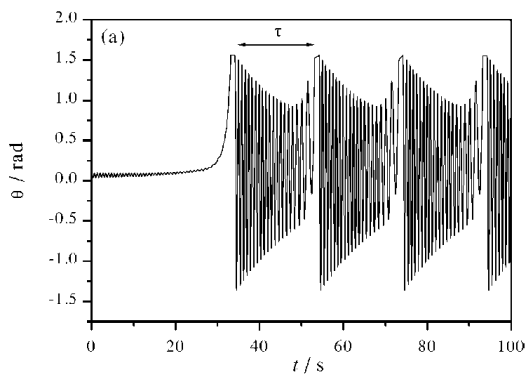
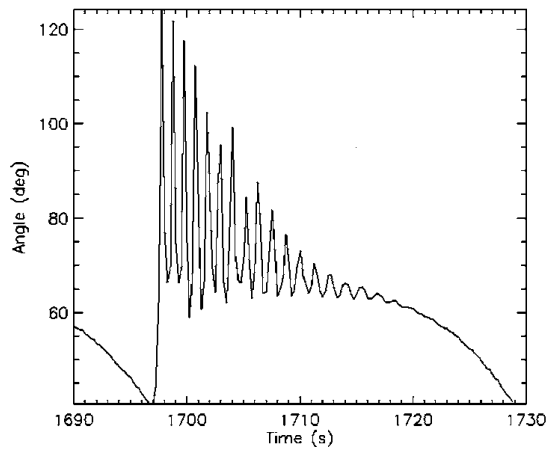
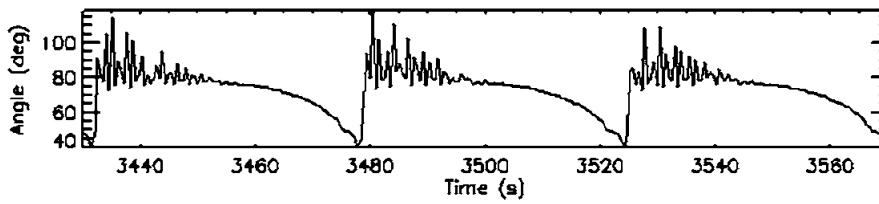


Figure 9.4: Guemez et al. Computational simulation of DB angle vs time.



(a)



(b)

Figure 9.5: Lorenz experimental measurement of one cycle in (a) and three cycles in (b) of the DB using electronic circuitry, with the angle taken with respect to the horizontal.

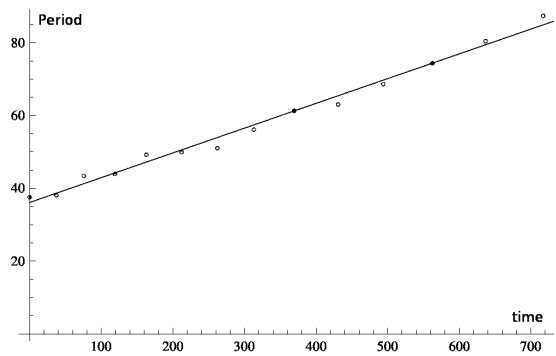


Figure 9.6: Our model: dip period vs time showing a strong linear relationship for $p = 0.002$; to be compared with figure 9.7.

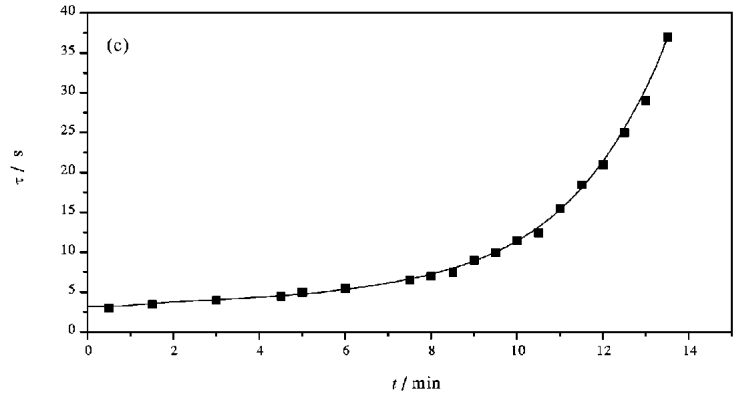


Figure 9.7: Guemez: dip period vs time measured in an enclosed chamber using a digital balance.

however, since for $p \ll 1$ we can approximate: $e^{p\Delta t} \approx 1 + p\Delta\tau$ and arrive at a linear relationship.

The equation of the line given by data analysis is $y = 36(1 + 0.002x)$ which fits very well to the linear approximation given. We would therefore assume that, over a longer time, we would have acquired the results in figure 9.7.

Phase Portrait

From the phase space diagram plotted in Figure 9.8 we get the most complete picture of the physics of our system. We can immediately see that it is a self-repeating cycle given that each cycle traces the same path in phase-space. We can also see that it is a dissipative system: the radius of the path decreases with each rotation while we can also clearly see the shift in equilibrium towards the later end of the cycle.

We can then conclude that our system should not conserve energy. Indeed, the physical DB does work and thus does not conserve energy itself.

Further Analysis

We now discuss the dissimilarities. The most noticeable difference is that of the gradient of the envelope enclosing our oscillations. This contrast, however, is quite minute and does little to diminish the quality of our model with respect to the actual physical process. This can obviously be attributed to the over-simplification of our model and is, arguably, a trivial disparity to encounter in light of the numerous successes.

We also note that our increasing period plot in figure 9.6 did not match that of figure 9.7. However this is not of great importance as figure 9.7 was measured in an isolated, controlled environment under specific conditions; we

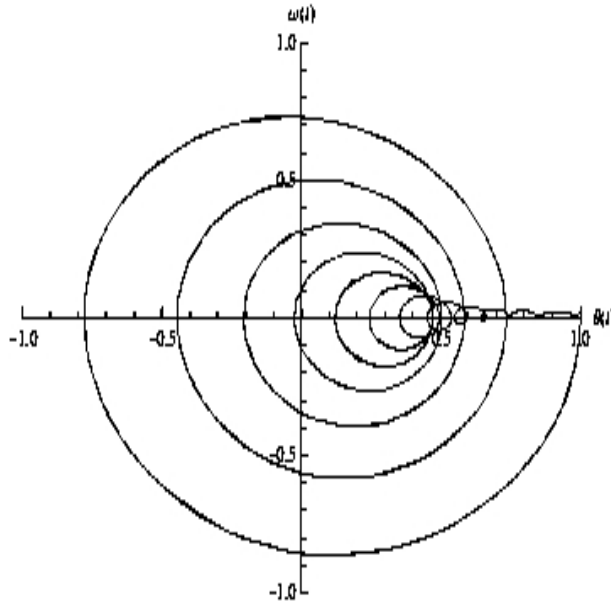


Figure 9.8: Phase portrait of DB for three cycles sans decaying force. We clearly see the stability of the model as the paths are all identical.

should therefore only concern ourselves with the general behaviour, which we have acquired.

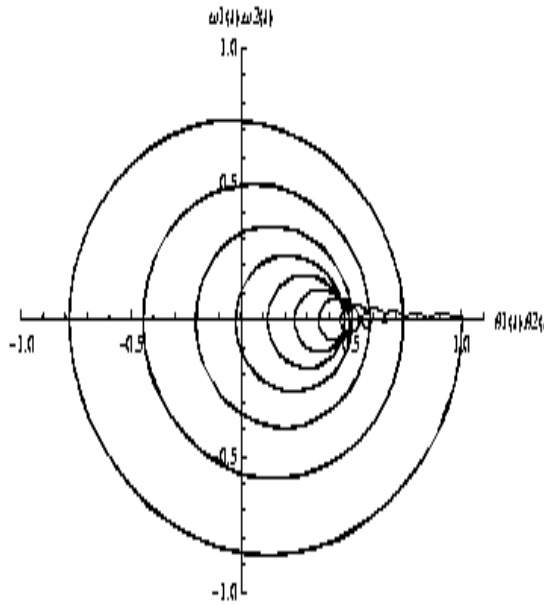
5. RESULTS: COUPLED DIPPY BIRDS

We now end our discussion of the results of Guemez and Lorenz, having established a satisfactory comparison, and discuss the unique results obtained by this project. Having assembled our model as discussed in section 3, we tested our model for simple in-phase and anti-phase cases to ensure we would achieve expected behaviour. Once satisfied we then began to vary α_1 in order to observe period doubling and/or tripling, and signs of chaos.

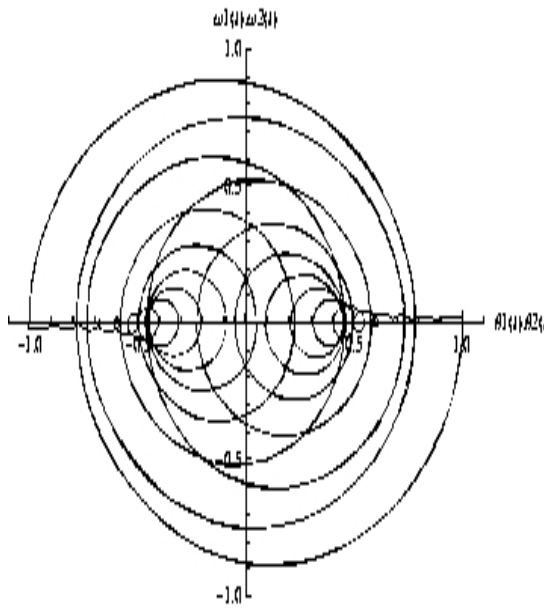
In-phase and Anti-phase

Since the focus on this part of the project was on deriving chaotic behaviour we set $p_1 = p_2 = 0$ for simplicity as we would not expect the force decay to be influential in the coupled case over short time periods. This also shortened our computing time. Hence both birds have identical conditions and from the overlayed phase-portrait in Figure 9.9a, we see the DBs' paths in phase-space are identical, which is the expected result.

We then set the DBs such that they should oscillate in anti-phase. The resultant over-layed phase-portrait in figure 9.9b clearly shows the anti-phase behaviour.



(a)

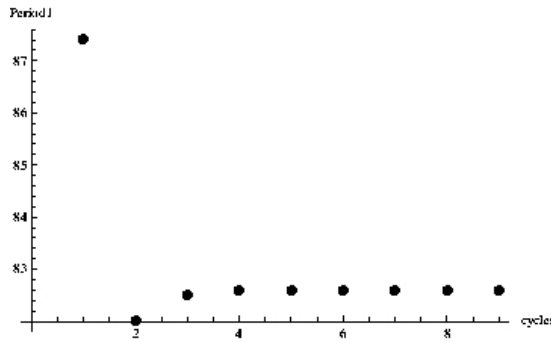


(b)

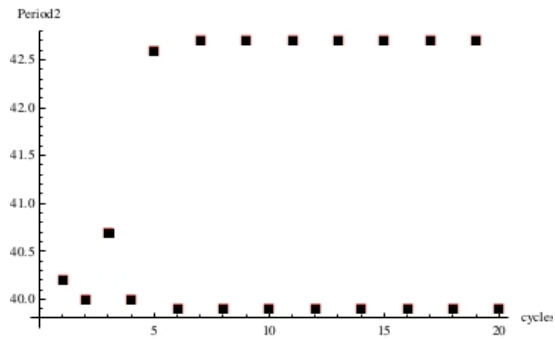
Figure 9.9: Overlayed phase portraits of birds 1 and 2. 9.9a) Birds are in phase, tracing the same path in phase-space and in 9.9b) in anti-phase for three cycles having paths out of phase by 180 degrees.

Period Doubling, Tripling

Having established a satisfactory model we wished to observe period doubling or tripling in bird 1 by varying α_2. We significantly reduced our step size to $j = 10^{-7}$ to increase our accuracy. Figures 9.10a and 9.10b show the periods of birds 1 and 2 respectively for each cycle. We can clearly see that after an initial settling period of a few cycles the birds assume a repeated pattern with bird 2 displaying period doubling.



(a) Bird1 Period

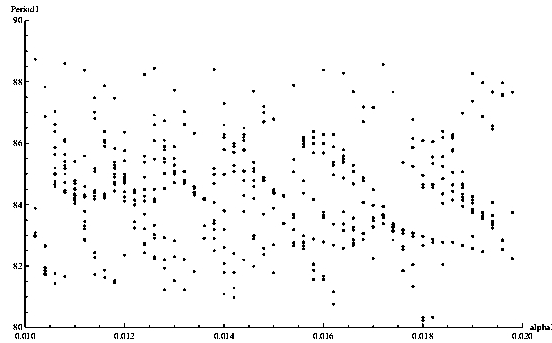


(b) Bird2 Period

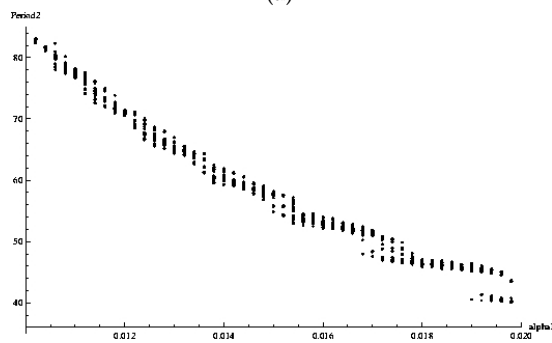
Figure 9.10: Periods of DBs demonstrating in 9.10a), regular behaviour for bird 1 and in 9.10b) period doubling for Bird 2, after some initial settling time.

CHAOS

Attempts were made at establishing bifurcation diagrams of period vs α_1 for both birds and observe chaotic behaviour. We varied α_1 from 0.01 to 1 in increments of 0.01. The diagrams obtained are given in Figures 9.11a and 9.11b. It is clear these are not typical bifurcation diagrams, but figure 9.11a does display hints of chaotic behaviour. However this is not a convincing example and more analysis is needed before we can say for sure. The apparent randomness of the period may indicate that chaotic behaviour is present but this may be a symptom of the aforementioned relaxing period.



(a)



(b)

Figure 9.11: Bifurcation Diagrams for Bird 1 in 9.11a, and Bird 2 in 9.11b) which show some signs of chaotic behaviour.

6. CONCLUSIONS AND DISCUSSION

Single Dippy Bird Model

Our single bird model was quite successful in replicating the general motion of a physical dippy bird and compared very well with the results of Lorenz [1] and Guemez [2]. We were able to devise and observe all three time scales thus completing our physical description. Given the simplicity of our model, we believe further work into fine-tuning it based on any physical observations would be extraneous and unnecessary as it provides a thorough and quantitative representation as it is.

Coupled Bird Model

For the coupled birds we validated our model by demonstrating the predicted in-phase and anti-phase motion and went further to observe period doubling for Bird 2. We do however note that period tripling was observed for simulations where the time-step was of order 10^{-3} but then collapsed into period doubling where in the *same* simulation the time-step was of order 10^{-7}. We

concede that the time-step in our observed period doubling may not have been fine enough to avoid this error.

Hints of chaotic behaviour were observed in our simulations, however we would like to investigate this further and determine it if it is indeed present.

Physical Model and Future Work

Although no future work will be done on this project, had we more time, we would have liked to investigate the effects of a non-linear force model for the single DB model. We would also have liked to test our model against a real DB and try to make some predictions as to its motion.

We would also have liked to re-run the simulations in the period doubling case with as small a time-step as possible to erase any doubt that it was indeed present, given that period tripling was seen to collapse when a smaller time-step was used.

More simulations would certainly have been carried out on the chaotic model to fully exhaust any avenues in which it might be observed. We would take the variable parameter over a longer range, or vary β_1 or γ_1 also to see their influence. However, we can not say at this time if chaotic behaviour will be determined for certain.

7. ACKNOWLEDGMENTS

With thanks to Dr. Stefan Hutzler and Dr. Patterson for the opportunity to work on this project and for their help in its completion, and to Mr. David Whyte for sharing his invaluable knowledge of *Mathematica* programming.

BIBLIOGRAPHY

[1] R. Lorenz, "Finite-time thermodynamics of an instrumented drinking bird toy", Am. J.Phys. **74**(8), 677-682 (2006).

[2] J. Guemez, R. Valiente, C. Fiolhais, and M. Fiolhais, "Experiments with the drinking bird", Am. J. Phys.**71**(12), 1257-1263 (2003).

THE CONTRIBUTORS

Daron Anderson is a student of mathematics at Trinity College Dublin entering his final year. [andersda@tcd.ie]

Kieran Cooney has completed two years of his undergraduate in mathematics and physics at University College Cork. He will spend his third year on exchange at the University of California, Berkeley. [111434868@umail.ucc.ie]

Brenden Daniel Williamson has an undergraduate degree in actuarial mathematics from Dublin City University. After an eight month internship at the Hamilton Institute of NUI Maynooth he decided to pursue postgraduate study, and will begin a PhD at Duke University, North Carolina, this September. [brendan.williamson3@mail.dcu.ie]

James Fennell has just completed an undergraduate degree in mathematical sciences at University College Cork. In September 2013 he begins the PhD program in mathematics at the Courant Institute of Mathematical Sciences, New York University. [jamespfennell@gmail.com]

Michael Hanrahan has completed two years of medicine at University College Cork. [111306766@umail.ucc.ie]

Thomas P. Leahy is entering his fourth and final year of financial mathematics and economics at the National University of Ireland, Galway. [t.leahy2@nuigalway.ie]

Anthony O'Farrell is professor of mathematics at the National University of Ireland, Maynooth. His article in this issue of the *IUMM* was first given as a talk at the Irish Mathematics Students Association conference in University College Dublin, March 2012. [anthonyg.ofarrell@gmail.com]

Anthony James McElwee has just completed an honours degree in electrical engineering at University College Dublin. He is currently considering postgraduate courses in simulation and mathematical modelling for September. [anthonyjames.mcelwee@gmail.com]

Glenn Moynihan has just finished an undergraduate degree in theoretical physics at Trinity College Dublin. He begins a PhD at Trinity this September in computational condensed matter. [glenn.moy@gmail.com]

Sean Murray has just finished an undergraduate degree in theoretical physics at Trinity College Dublin where he will stay to study for a masters in high performance computing. [smurray4@tcd.ie]

MAKING THE IUMM

The Summer 2013 issue of the *Irish Undergraduate Mathematical Magazine* was typeset using Michael Spivak's *MathTime Professional 2* typeface for mathematical formulae and *TEX Gyre Pagella* for all other text. A free "lite" version of the former can be downloaded at `http://bit.ly/mtpro2`. The latter comes in the standard LaTeX package `tgpagella`.

In combining typefaces from two different families we committed a typographical sin (observe the difference between a and a), but we think the results are simply too pretty to mind (and Mr. Spivak agrees).

◇

A number of figures were redrawn in vector graphic form for increased clarity. Those, for example, on pages 31 and 49 were drawn using PGF/tikz, a LaTeX extension in which figures are described in text form using a LaTeX-like syntax. The results are highly scalable and generally more visually pleasing than alternative figure generation methods.

The open-source vector-graphics program *Inkscape* provides an easy-to-use GUI alternative, and was used to draw the figures on pages 66 and 74, as well as the cover.

◇

The design of the magazine was guided by that of the *American Mathematical Monthly*. The magazine was constructed using a custom-made LaTeX journal system. Each article and its associated files are kept in separate sub-directories and included in the magazine through a one-line LaTeX command that automatically handles titling, headering, and the reseting of figure and section counters. Those interested in typesetting can contact the editor for a (free) copy of the software.